ROUTING AND NETWORK MOBILITY MANAGEMENT IN NEXT GENERATION SATELLITE NETWORKS

by

Ömer Korçak

B.S. in Computer Engineering, Boğaziçi University, 2002

M.S. in Computer Engineering, Boğaziçi University, 2004

Submitted to the Institute of Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2009

ROUTING AND NETWORK MOBILITY MANAGEMENT IN NEXT GENERATION SATELLITE NETWORKS

APPROVED BY:

Assoc. Prof. Fatih Alagöz (Thesis Supervisor)	……
Prof. Emin Anarım	……
Prof. M. Ufuk Çağlayan	……
Assist. Prof. H. Ali Mantar	……
Assoc. Prof. Tuna Tuğcu	……

DATE OF APPROVAL: 15.05.2009

ACKNOWLEDGEMENTS

This thesis originates from my work at the Satellite Networks Research Laboratory (SATLAB) of Computer Engineering Department in Boğaziçi University. During this time, many people have contributed in various ways to the successful outcome of this work.

Firstly, I am grateful to my thesis supervisor, Assoc. Prof. Fatih Alagöz for his friendly behavior, clever guidance, genial talks and useful comments during my thesis. I have been really fortunate to have him as my advisor. It was satisfactory and enjoyable to work with him. I also thank to Assoc. Prof. Tuna Tugcu, Prof. Emin Anarım, Prof. M. Ufuk Çağlayan, and Assist. Prof. Hacı Ali Mantar for their participation to my thesis jury, and for their interest, concern, and very useful comments.

I am greatly obliged to my parents, and the rest of my family. They continuously support me in all respects during my whole life, and they were being there for me whenever I needed them. They have a major contribution on this work. Special thanks to my self-sacrificing and compassionate wife. She is with me in all respects, and I cannot imagine better companionship than she does.

I would like to thank to all of my faithful and sincere friends. Without their support, life would become unbearable. Special thanks to Osman Mülayim, Onur Avcıoğlu and Toygun Yaprakçı for their continuous aids and for being such good friends.

I gratefully acknowledge the financial support of TUBITAK under National PhD scholarship, and the financial support of DPT under grant numbers DPT-03K 120250 and DPT-2007K 120610.

Last, but not least, I thank to all my lab-mates for their valuable assistance, responsiveness, technical help, and mainly for having contributed to this long journey.

ABSTRACT

ROUTING AND NETWORK MOBILITY MANAGEMENT IN NEXT GENERATION SATELLITE NETWORKS

Satellite networks are an attractive option to provide broadband telecommunication services to globally scattered users, due to their extensive geographic coverage, high bandwidth availability, inherent broadcast capabilities, etc. Satellites rotating in geostationary orbit (GEO) are very well suited for broadcast services, but they suffer from high free space attenuation and long delays. On the contrary non-geostationary (NGEO) systems consisting of Medium Earth Orbit (MEO) and Low Earth Orbit (LEO) satellites offer smaller latency, lower free space loss, and better re-use of available ground-space communication frequencies, hence they are more suitable for most applications (especially for those running in real-time). However, these advantages come with a price: Footprints of satellites at lower altitudes are smaller, and global coverage can be provided by higher number of satellites that are connected each other with inter-satellite links (ISL). Moreover, lower orbit satellites move with higher speeds relative to the Earth's surface, resulting in high dynamic in the network topology. Dynamics of the satellite constellation constitute major challenge in providing efficient routing and quality of service (QoS) for rapidly-growing real-time multimedia services. On the other hand, regular NGEO satellite networks has some facilitating features like periodicity, predictability and having highly symmetric and regular topology. For efficient networking in NGEO satellite networks, all these features should be considered.

In this thesis, we clarify features of satellite systems that differ them from their terrestrial counterparts and propose novel routing and network mobility management techniques in NGEO satellite networks. Firstly, we make use of geometrical properties of the network topology, and propose a priority-based adaptive routing (PAR) algorithm. Next, we focus on handling the mobility of network by utilizing satellites with Earth-fixed footprints, and extend a well-known mobility handling technique called Virtual Node

(VN). We propose Multi-state Virtual Network (MSVN) topology that alleviates deficiencies of VN concept. We clarify potential advantages of MSVN by developing efficient handover mechanisms and beam management techniques in MSVN-based satellite systems. Finally, we investigate efficient integration of NGEO satellites with High Altitude Platforms (HAPs) via high-capacity free-space optical links for carrying dense and real-time multimedia traffic. Considering the mobility and resource limitations of satellites, we propose an efficient solution for the optimal link establishment problem between HAPs and satellites.

ÖZET

GELECEK NESİL UYDU AĞLARINDA YOL ATAMA VE AĞ HAREKETLİLİĞİNİN YÖNETİMİ

Telekomünikasyon endüstrisinin hızlı bir küreselleşme sürecine girmesi ile, geniş coğrafi kapsama alanı ve çoğa gönderim kabiliyetleri gibi önemli özellikleri bulunan uydu sistemlerinin iletişim alanındaki rolü giderek artmaktadır. Yerdurağan-yörüngedeki (GEO) bireysel uyduların kullanılmasıyla başlayan uydu iletişimi, bu uyduların yerden uzaklığından kaynaklanan yayılım gecikmesi gibi nedenlerden dolayı gerçek-zamanlı ve interaktif uygulamalar için elverişli değildir. Dolayısıyla son yıllarda uydu piyasasında orta-yörünge (MEO) ve bilhassa alçak yörünge (LEO) uydu sistemlerine yönelmeler olmuştur. Yerdurağan olmayan (NGEO) bu uydu sistemlerinin düşük yayılım gecikmesi, düşük sinyal kaybı ve frekansların daha verimli kullanılabilmesi gibi avantajları vardır. Fakat bu avantajlar, bir takım zorlukları da beraberinde getirir. LEO uydularının kapsama alanlarının GEO uydularına kıyasla az olması nedeniyle, global kapsama için birbirleriyle iletişim bağları olan çok sayıda uydu gerekmektedir. Ayrıca, alçak yörünge uydularının yere göre yüksek bağıl hızları, uydu ağ topolojisinin hareketli olmasına neden olur. Uydu ağının hereketliliği, verimli yol atama ve servis kalitesini sağlamak için en önemli problemi teşkil etmektedir. Diğer taraftan, uydu ağlarının simetrik ve düzgün bir yapıya sahip olması, ağ hareketliliğinin önceden tahmin edilebilir ve periyodik olması gibi bir takım özellikleri vardır. Uydu ağları üzerinden verimli bir iletişim sağlamak için bunlar gibi bütün özelliklerin hesaba katılması gerekmektedir.

Bu tezde, öncelikle uydu sistemlerini yer ağlarından ayıran temel özellikleri, bu özelliklerin doğurduğu ihtiyaçları ve getirdiği sonuçları sınıflandırarak ortaya çıkardık. Ardından, yeni yol atama ve ağ hareketliliğinin yönetim teknikleri önerdik. Birinci olarak, ağ yapısının geometrik özelliklerinden faydalanarak öncelik tabanlı uyarlamalı bir yol atama tekniği geliştirdik. İkinci olarak, ağ yapısının hareketliliğinin üstesinden gelmek için yersabit ayak izli uydu sistemlerini ele aldık. Bu sistemlerde hareketliliği yönetmek için

çok bilinen bir yöntem olan sanal düğüm (VN) tekniğininin eksiklerini telafi eden çok durumlu bir sanal ağ mimarisi (MSVN) önerdik. MSVN tabanlı uydu sistemlerinde verimli el değiştirme ve ışın huzmelerinin yönlendirilmesi teknikleri geliştirerek önerilen mimarinin olası avantajlarını ortaya koyduk. Son olarak, yerdurağan olmayan uydular ile yüksek platformların (HAP) yüksek kapasiteli optik bağlar kullanarak verimli entegrasyonu problemine yöneldik. Uyduların hareketliliğini ve kaynak limitlerini göz önünde bulundurarak hangi uydular ile hangi HAP'lar arasında optik bağ kurulacağına karar verme problemine çözüm getirdik.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

A	Aggregate traffic generated worldwide
A_{avg}	Average of elevation angles
A_{avg}^{t}	Average of elevation angles at time t
a_h	Activity percentage in hour h
B / λ	Link bandwidth
$B_{\max}$	Maximum amount of onboard buffering
B_{DEM}^{f}	Bandwidth demand of footprint f
$B_{\text{DIR}}^{f_2,f_1}$	Amount of bandwidth directed from footprint f_2 to footprint f_1
$B_{\text{DIR}}^{\max}$	Maximum possible amount of directed bandwidth
B_{STS}^{f}	Satisfied amount of demand for footprint f
B_{SUP}	Average bandwidth supplied to a footprint area
$B_{\text{SUP}}^{\text{sat}}$	Bandwidth supplied by a satellite
DS	Dummy source
DT	Dummy sink
d	Threshold for number of deflections
$d_{\max}$	Angular distance to regular switching point
dir_{x}	East or west
dir_{y}	South or north
$dist(x,y,t,k)$	Distance between zone $Z_{x,y}$ and zone $Z_{t,k}$
$E_{N\times M}^{t}$	Existency matrix
$EA_{N\times M}^{t}$	Matrix of elevation angles
F_{e}	Existing aggregate flow in a link
f_{mn}	flow between m and n
$G(\gamma)$	Gain function
$H_{\max}$	Maximum number of HAPs a satellite can serve
$H_{\max}^{i}$	Maximum number of HAPs satellite i can serve
h	Current local hour

h_C	Number of hosts in continent C
h_H	Height of HAP
h_i	Number of hops packet p_i traversed
h_S	Height of satellite
$h_{x,y}$	Host density level of zone $Z_{x,y}$
$L(\gamma)$	Loss function
$L_{\text{E-E}}$	End-to-end handover latency
L_{FP}	Angular length of an Earth footprint
L_{ISL}	Latency in inter-satellite links
L_{SAT}	Inter-satellite angular distance
$L_{\text{UDL\_U}}$	Latency in uplink UDL
$L_{\text{UDL\_D}}$	Latency in downlink UDL
LD	Average link duration time
l_{q}	Average queue length
M	Number of HAPs
M_{FP}	Number of consecutive footprints served by fixed number of satellites
M_{FP}^{H}	Number of footprints in high service mode
M_{FP}^{L}	Number of footprints in low service mode
M_{S}	Total number of served HAPs
M_{SAT}	Number of satellites serving M_{FP} footprints
M_{U}	Total number of unserved HAPs
$mhd_{x,y}$	Minimum hop distance between satellite x and satellite y
N	Total number of satellites in a satellite network
N_{FP}	Number of footprint areas served by satellites along an orbit
N_{P}	Number of orbit planes
N_{S} , N_{SAT}	Number of satellites per orbit plane
N_{ST}	Number of states
$N_{\text{S/F}}$	Number of satellites per footprint area
$N_{\text{S/F}}^{\text{H}}$	Number of satellites per footprint area in high service mode
$N_{\text{S/F}}^{\text{L}}$	Number of satellites per footprint area in low service mode
$N_{\text{S/F}}^{\text{avg}}$	Average number of satellites per footprint area

n_{sd}	n_t - n_t^{sd}
n_t	Successfully transmitted data per second
n_t^{sd}	Amount of transmitted packets corresponding to *s-d* route
R_E	Radius of the Earth
r	Number of MSVSNs for an orbit
r_{max}	Ratio of d_{max} to L_{FP}
$S_{p,s}$	Satellite representation
$T(x,y,t,k)$	Traffic requirement from zone $Z_{x,y}$ to zone $Z_{t,k}$
T_{agg}	Aggregate demand
T_E	Self-rotation period of the Earth
T_{ISL}	Intra-orbit ISL link delay
T_S	System period
T_{sat}	Rotation period of a satellite
T_{SW}	Switching time
T_{UDL}	UDL link delay
Δt	Period of repeating optimization process
t_a	Length of the aging period
t_{a1}	Duration of active 1 state
t_{a2}	Duration of active 2 state
t_{a3}	Duration of active 3 state
t_F	Transmitted portion of F_e
t_{p1}	Duration of passive 1 state
t_{p2}	Duration of passive 2 state
Δt_{ST}	State interval
U_H	Utilization of HAPs
U_H^t	Utilization of HAPs at time t
$u_{x,y}$	User density of zone $Z_{x,y}$
$V_{N\times M}^t$	Visibility matrix
X / λ	Rate of F_e
$Z_{x,y}$	Terrestrial zone representation
α_I	Inclination angle of satellite orbit
α , β	Priority metric parameters

γ	Variable used for favoring existing links
Σ	A large negative number
ε_{SH}	Elevation angle between a satellite *S* and a HAP *H*
$\varepsilon_{\min}$	Minimum elevation angle
η	Normalization factor
λ	Rate of poisson process
$\lambda_{p,s}(t)$	Latitude of satellite $S_{p,s}$
μ	Priority metric
μ_{sd}	Priority metric for traffic traversing on *s*-*d* route
μ^{o}, μ^{n}	Temporary variables used in aging mechanism
$\varphi_{p,s}(t)$	Longitude of satellite $S_{p,s}$
$\Delta\phi$	Phase difference between adjacent orbits
ω_{E}	Angular speed of the Earth
ω_{S} , ω_{SAT}	Angular speed of a satellite

ACO	Ant Colony Optimization
AS	Autonomous System
ATM	Asynchronous Transfer Mode
BGP	Border Gateway Protocol
B-ISDN	Broadband Integrated Services Digital Network
CDMA	Code Division Multiple Access
DEFAR	Deflection Enabled FAR
DEPAR	Deflection Enabled PAR
DRA	Datagram Routing Algorithm
DVMRP	Distance Vector Multicast Routing Protocol
DVTR	Dynamic Virtual Topology Routing
E	East
ELB	Explicit Load Balancing
ePAR	Enhanced PAR
FAR	Fixed Adaptive Routing
FD	Flow Deviation
FP	Footprint

FSA	Finite State Automaton
gcd	greatest common divisor
GEO	Geostationary Earth Orbit
GMT	Greenwich Mean Time
GRI	Global Routing Info
GT	Ground Terminal
HAP	High Altitude Platform
ILP	Integer Linear Programming
IP	Internet Protocol
ISL	Inter-satellite Link
IOL	Inter-orbit Link
lcm	least common multiple
LEO	Low Earth Orbit
LER	Label Edge Router
LRI	Local Routing Info
LSP	Label Switching Path
MCC	Mission Control Center
MEO	Medium Earth Orbit
MFMR	Maximum Flow Maximum Residual
MPLS	Multi-protocol Label Switching
MOSPF	Multicast Routing Extensions for OSPF
MSVN	Multi-state Virtual Network
MSVN-SHO	MSVN-based Soft Hand-over
MSVN-SSHO	MSVN-based Semi-soft Hand-over
MSVSN	Multi-state Virtual Sub-network
MW	Microwave
MWMC	Maximum Weighted Maximum Cardinality
N	North
NGEO	Non-geostationary Earth Orbit
NGF	Net Gain Function
OBP	On-board Processing
OPW	Oldest Packet Win
OSPF	Open Shortest Path First

PAR	Priority-based adaptive Routing
PAR-MD	PAR for Minimum Delay Path Routing
pdf	probability density function
PRP	Probabilistic Routing Protocol
RAR	Random Adaptive Routing
RIP	Routing Information Protocol
RPW	Random Packet Win
RVM	Reverse-path Multicast
S	South
SGRP	Satellite Grouping and Routing Protocol
SHW	Shortest Hop Win
QoS	Quality of Service
TCD	Traffic Class Dependent
UDL	Up Down Link
VN	Virtual Node
VN-HO	Virtual Node Hand-over
VoIP	Voice over IP
W	West
WDM	Wavelength Division Multiplexing

1. INTRODUCTION

1.1. Satellite Systems Overview

With the rapid globalization of the telecommunications industry, satellites are expected to widely appear in future telecommunication systems, due to their potential advantages over terrestrial networks. Firstly, due to their extensive geographic coverage, they are able to provide services over a wide geographical area, including remote, rural, urban and inaccessible areas. They have a global reach with very flexible bandwidth-on-demand capabilities. Moreover, with their coverage superiority, satellites represent the most attractive solution for broadcast and multicast services which constitutes huge portion of services offered by next generation wireless systems. They are also beneficial due to their flexible deployment feature and good support for mission-critical applications. Satellites can be used also as a safety valves for terrestrial networks, so that network failures (e.g. due to environmental disasters) or congestion problems can be easily recovered. For these reasons and more, satellite systems are seen as an attractive solution to realize the global telecommunications infrastructure.

1.1.1. Basic Satellite Orbits

Internetworking with satellites began successfully with the use of individual satellites rotating at geostationary Earth orbit (GEO). GEO satellites are located at approximately 35,786 km above the Equator. The angular velocity of the satellite in this orbit is equal to the angular velocity of the Earth's self-rotation, hence the satellite appears stationary when observed from the surface of the Earth. GEO satellites can serve very large areas. Three or four satellites are sufficient for covering majority of the Earth. However coverage of high latitudes is impossible above 81° latitude and rarely possible above 75°, so full Earth coverage cannot be achieved by using any purely GEO constellation [1].

Due to their stationary and large coverage areas, GEO satellites are well suited for broadcast services. For more than three decades, they have been successful in providing

commercial services such as direct video broadcasting. However, GEO satellites suffer from high free space attenuation and long delays. Typical value of one-way end-to-end propagation delay between two ground terminals via a GEO satellite is around 270 ms, which is undesirable for most emerging services, especially for interactive and real-time applications. Moreover GEO satellites face limitations on the minimum cell size projected on the Earth's surface. Therefore focus has been directed towards development of lower orbiting non-geostationary (NGEO) systems.

Among possible orbit selections, medium Earth orbit (MEO) satellites with an altitude between 5,000 and 13,000 km, and low Earth orbit (LEO) satellites between 500 and 1,500 km were considered [2]. These altitude selections assure that the satellites reside outside the two Van Allen belts to avoid the radiation damage to electronic components installed in satellites. Typical one way end-to-end propagation delay for MEO satellites range between 110 – 130 ms. For global coverage of Earth, 10 – 15 satellites are needed. MEO satellites appear in motion when observed from the Earth, and they rotate around the Earth in approximately 4 – 6 hours. Visibility time of a MEO satellite to a ground station is in order of tens of minutes before handover must take place. ICO system [3] is an example of communications satellites operating at MEO level.

For a LEO satellite, one way end-to-end propagation delay is typically around 20 – 25ms, which is comparable to that of a terrestrial link. Since they are closer to the Earth's surface, the necessary antenna size and transmission power level are much smaller. On the other hand, LEO satellites move rapidly with an orbit period of around 2 hours, which necessitates frequent handovers. Moreover, coverage areas of satellites are smaller, and more than 32 satellites are required to provide global coverage. The actual number of satellites used in a LEO constellation depends upon the altitude of the orbit, coverage required and the minimum elevation angle desired for communication. LEO satellite constellations with large number of satellite with small coverage areas are more complex systems than GEO satellite networks, but they offer larger system capacities by providing larger amount of frequency reuse. Globalstar [4], Iridium [5], and Teledesic [6] systems are examples of LEO communications satellites.

1.1.2. Basic Satellite Constellation Topologies

A regular NGEO satellite constellation is characterized by a number of system parameters: Number and altitude of satellites, number of orbits and satellites per orbit, how to deploy the orbits, and how to inter-connect the satellites.

According to the way the orbits are deployed, different types of constellations are obtained. If all orbit planes are deployed along a semi-circle when viewed from a pole, resulting constellation is a so-called *π-constellation* [7]. In π-constellations, there are two extreme orbits which are adjacent, but whose satellites move in opposite directions. As a result, a *seam* appears between these two orbit planes and potential inter-plane ISLs passing over the seam must hand-over frequently. Moreover, usually π-constellations suffer from extensive polar coverage, and do not provide dual satellite visibility in low latitudes. In order to avoid this kind of problems *2π-constellations* are proposed. In 2π-constellations, ascending nodes are equally spaced along the full 360˚ of the equatorial plane.

Another important parameter of a satellite constellation is the orbital inclination angle, which is defined as the angle between the orbit plane and the equatorial plane. If the inclination angle of an orbit is closed to 90˚, satellites pass over the polar regions and the orbit is called *polar orbit*. Usually, π-constellations use polar orbits for coverage reasons, and they are called *polar constellations* [7]. (Note that they are also named as Walker star constellations [8], since orbital pattern appears as a star '*' in polar view). On the other hand, inclined orbits (orbits with low inclination angles) are better suited for 2π-constellations. 2π-constellations with inclined orbits are shortly called *inclined constellations* (or Walker delta constellations [8], since orbital pattern appears as a delta 'Δ' in polar view). While Iridium system and Teledesic system design have polar constellation topology, Globalstar and ICO systems are inclined constellations. System characteristics of these major NGEO satellite systems are summarized in Table 1.1.

If the satellites in adjacent co-rotating orbits have same latitudes, then *phase difference ($\Delta\phi$) between adjacent orbits* is said to be zero. However, the satellites in

adjacent orbits may be shifted relative to each other to provide coverage without gaps. This is more important in polar constellations. In Iridium system, $\Delta\phi$ is equal to π/N_S, where N_S is the number of satellites in an orbit. On the other hand, in Globalstar, $\Delta\phi$ is smaller and is equal to $\pi/(N_P \cdot N_S)$, where N_P is the number of orbit planes. In ICO system there is no phase shift between two orbits.

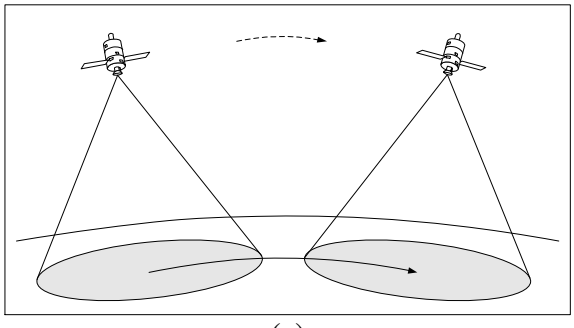
Initial NGEO satellite constellations (such as Globalstar) do not include on-board processing (OBP) and inter-satellite links (ISLs). In such systems, a satellite must have a gateway station in view to provide service to any users it may see, due to the lack of inter-satellite linking. However, if there are no gateway stations to cover certain remote areas, service cannot be provided in these remote areas, even if the satellites may fly over them. Therefore, for the better utilization of satellites and to increase the performance of the system, new NGEO systems usually present OBP capabilities, including modulation/remodulation, decoding/recoding, transponder/beam switching and routing. In such systems, any two satellites within line-of-sight can be connected to each other via ISLs. In Iridium system, each satellite can be linked to its four neighbors (if its neighbor is not co-rotating). In Teledesic system concept, there exists up to eight ISLs per satellite. The use of ISLs raises the issue of routing in the satellite network, which we will investigate in detail in this thesis.

Table 1.1. Characteristics of some major NGEO satellite constellations

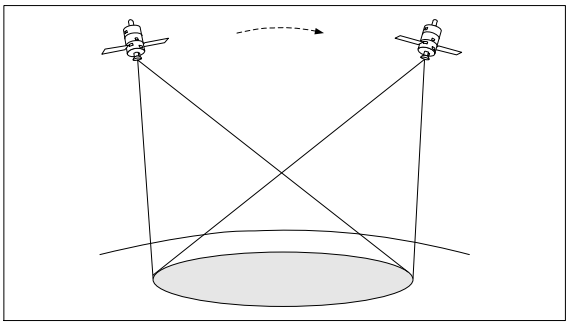
	ICO	**Globalstar**	**Iridium**	**Teledesic**
Orbit type	MEO	LEO	LEO	LEO
Constellation type	inclined	inclined	polar	polar
Altitude (km)	10,355	1,410	780	1,375
Number of satellites	10	48	66	288
Number of orbits	2	8	6	12
Orbit period (min)	358.9	114	100.1	98.8
Inclination angle	45˚	52˚	86.4˚	84.7˚
Satellite visibility time (min.)	115.6	16.4	11.1	2.32
Minimum elevation angle	10˚	10˚	8.2˚	40˚
Beam per satellite	163	16	48	64
Satellite antenna	fixed	fixed	fixed	steerable
Footprint diameter (km)	12,900	5,850	4,700	1,412
OBP	No	No	Yes	Yes
ISL	No	No	Yes	Yes
Coverage	global	within ±70˚ latitude	global	global

1.1.3. Earth Coverage

Coverage area or *footprint* of a satellite is the union of the areas (*cells*) covered by the spot beams of that satellite. If the satellite has fixed antenna system and it doesn't steer its beams, then footprint of the satellite sweeps across the Earth's surface with a constant velocity, as shown in Figure 1.1(a). This type of satellite systems have *satellite-fixed (or nadir pointing) footprints*. On the other hand, if the satellite is capable of steering its beams, coverage of the spot beams can be fixed for a time duration. In this case, cells covered by the spot beams are said to be *Earth-fixed*. The satellite can make up of its motion by steering all of its beams simultaneously and switching them synchronously, which leads to *Earth-fixed footprints* as shown in Figure 1.1(b). After some time, all the satellites will be moving away from their corresponding footprints, and the system periodically reassigns each satellite to a new fixed footprint. Earth-fixed footprint concept is introduced in [9]. Although this technique comes with the cost of degradation in the elevation mask used in the system (or increase in the number of satellites), it has the potential to simplify the handovers as we will describe in Section 5.2. Teledesic system concept supports Earth-fixed cells.



(a)



(b)

Figure 1.1. (a) Satellite-fixed and (b) Earth-fixed footprints

1.2. Research Objectives and Solutions

In this thesis, new and efficient routing and network mobility management mechanisms are proposed to address challenges in emerging NGEO satellite networks. First, we provide a thorough survey which is an extension of the work that we presented in [10]. Then, we investigate the following three areas under this research:

1.2.1. Distributed Routing in NGEO Satellite Networks

In a NGEO satellite constellation with ISLs, there could be many shortest paths between two satellites in terms of hop count. An efficient routing algorithm should effectively use these paths in order to distribute traffic to ISLs in a balanced way and to improve the performance of the system. In order to achieve this goal, we propose and evaluate a novel priority-based adaptive shortest path routing (PAR) scheme [11, 12]. PAR algorithm sets the path towards the destination in a distributed manner, using a priority mechanism depending on the past utilization and buffering information of the ISLs. We also make some extensions to the proposed algorithm, in order to further increase the routing performance [12, 13].

1.2.2. Network Mobility Management in Earth-fixed NGEO Satellite Systems

Handling network mobility in a highly dynamic LEO satellite network is a critical issue to achieve seamless and efficient integration of satellite and terrestrial networks. In Earth-fixed satellite systems, this task could be simplified by representing the network with a more static virtual topology. Virtual node (VN) approach is widely explored in satellite networks research; however, it has some deficiencies due to necessity of one-to-one correspondence between virtual nodes and physical satellites. We propose a generic virtual topology model, namely multi-state virtual network (MSVN) architecture that alleviates these deficiencies. A new mathematical model for MSVN is introduced along with its potential contribution to the overall system availability [14, 15]. Furthermore, possible handover mechanisms in Earth-fixed satellite systems are investigated, and new efficient handover mechanisms for both VN-based systems and MSVN-based systems are proposed

and compared [15, 16]. Finally, we also deal with management of satellite beams in order to achieve optimal system availability and performance in Earth-fixed NGEO satellite systems [17].

1.2.3. Integration of NGEO Satellite Systems with High Altitude Platforms (HAPs)

NGEO mobile satellite systems integrated with HAPs may have great potential in the next generation telecommunication services. Efficient integration of NGEO satellite network and HAPs is important for maximizing availability and performance of the system. In this context, we focus on establishment of high-capacity free-space optical links between HAPs and mobile satellites with limited resources. We formulate an optimization problem for matching HAPs and satellites in such a way that the utilization of HAPs is maximized together with the average elevation angle between HAPs and satellites. We also propose a method to avoid frequent switching of optical links. We come up with a polynomial-time solution approach for the formulated problem, demonstrate numerical results in sample scenarios for various system parameters, and discuss the reasonable selection of these variables [18].

1.3. Outline of the Thesis

Conventional networking mechanisms (such as routing protocols) proposed for terrestrial networks are not directly applicable to satellite networks due to some features particular to satellite systems. Mobility of satellites constitutes major challenge in providing efficient routing and quality of service (QoS) for rapidly-growing real-time multimedia services. Moreover, satellite nodes and satellite constellation topologies present some specific properties that should be taken into account for efficient networking. Chapter 2 provides a brief background on features of NGEO satellite networks. Next, Chapter 3 surveys various routing and network mobility management techniques proposed for satellite networks. Several technical issues regarding the application of these techniques are discussed. Chapter 4 presents and evaluates *priority-based adaptive routing algorithm*, which is a proactive distributed routing algorithm proposed for NGEO satellite networks. Chapter 5 describes network mobility management issues in Earth-fixed NGEO satellite systems and presents a novel mobility modeling called *multi-state virtual network*

architecture. In the scope of Earth-fixed NGEO satellite systems, novel handover mechanisms and efficient beam management techniques are proposed and discussed. Chapter 6 focuses on efficient integration of NGEO satellites with HAPs, using free space optical links. Considering the mobility and resource limitations of satellites, an optimal assignment method of HAPs to satellites is proposed and evaluated. Finally, Chapter 7 summarizes the main advantages and contributions of the work presented in this thesis.

2. FEATURES OF NGEO SATELLITE NETWORKS

A conventional satellite constellation consists of a number of orbits at a certain altitude and with a given inclination angle, a number of satellites per orbit (plane) and ISLs between some satellite pairs. Keeping the orbit lower is more attractive, since it allows the reduction of the necessary antenna size as well as transmission power level and leads to lower communication delay. However, these advantages come with a price: Firstly, footprints of satellites at lower altitudes are smaller. Therefore, a higher number of satellites is needed for global coverage. In addition, lower orbit satellites move with higher speeds relative to the Earth's surface, resulting in high dynamism in the satellite constellation topology. The mobility of satellites constitutes major differences between satellite networks and their terrestrial counterparts. Moreover, satellite nodes have different capabilities than terrestrial nodes, and satellite constellation topologies present some specific geometric properties. In this chapter, we give a brief overview of these characteristics of satellite constellation networks. For high-performance networking, the effects of all these features should be considered.

2.1. Effects of Satellite Mobility

2.1.1. Dynamic Network Topology

While the ISLs between satellites in the same plane (intra-plane ISLs) are fixed in length, the length of the ISLs between satellites from different planes (inter-plane ISLs) changes depending on the movement of the satellites. In polar constellations, for example, intra-plane ISLs are the longest when satellites are over the equator and the shortest when they are over the polar region boundaries. Moreover, network connectivity can also vary. ISL connectivity between satellites may change based on the distance and the viewing angle between them. ISLs passing over seam are also switched on and off continuously. On the other hand, when a satellite enters a polar region, its adjacent inter-plane ISLs are switched off.

2.1.2. Predictability and Periodicity of Network Topology Changes

Although the topology of a satellite network rapidly changes, it is deterministic and can be predicted quite accurately. Moreover, the complete topology dynamics are periodic, i.e. it repeats itself with a known period.

2.1.3. Highly Dynamic and Non-homogeneous Traffic

Since satellites cover smaller areas in low orbit systems, traffic requirements are unbalanced due to the varying population density, which is high in the cities, low at rural areas and almost zero over the oceans, which cover 71 per cent of the Earth's surface. As satellites move, traffic received from terrestrial nodes varies continuously depending on the user density in the footprint area.

2.1.4. Necessity of Handovers

In connection-oriented satellite network structures, where ISLs are switched off due to mobility of the system, a link handover process is needed to maintain active connections. Link handover could be required either when ISL connectivity changes or when the link between a user node and a satellite becomes unavailable. Handovers of active communications between user terminals and satellites should be controlled considering the capabilities of the satellite antenna system. There are two general techniques: Asynchronous handover and synchronous handover, which are described in Section 5.2.

2.2. Limitations and Capabilities of Satellites

2.2.1. Limited Power and On-board Processing Capability

As more complex processing is done on satellites, they consume more power and their lifetime becomes shorter. There is a trade-off between the lifetime and the processing capability of satellites. Actually, a very long lifetime is not needed, since technology

improves very fast. One drawback of the satellites when compared to the terrestrial nodes is that once a satellite is launched, it is infeasible to upgrade its technology or to extend its storage and processing capabilities.

2.2.2. Difficulty of Implementing the State-of-the-art Technology

The long lead times between the design, development, launching and service stages of satellite systems usually make it difficult to implement the *state-of-the-art* technology. Therefore, satellites should not be designed specifically for use with just the current technology and interfaces with terrestrial networks should also be designed with a similar approach.

2.2.3. Broadcast Nature of Satellites

They offer great potential for multimedia applications with their ability to broadcast and multicast large amount of data over a very large area.

2.3. Features due to the Nature of Satellite Constellations

2.3.1. Higher Propagation Delays

Because of the long distances between satellites and the high altitude of the constellation, propagation delay can be considered the most important cost factor in satellite networks. As the altitude of the orbit increases, its effect becomes more evident.

2.3.2. Fixed Number of Nodes

Disregarding the satellite failures, generally the number of nodes in a satellite network is fixed unlike most terrestrial networks, where new links or destinations can be added on a daily basis.

2.3.3. Highly Symmetric and Uniform Structure

Since the constellation topology is highly symmetric and uniform, there can be many alternate paths between two satellite nodes. Selection of the most appropriate path can effectively increase the utilization of the system.

3. ROUTING AND NETWORK MOBILITY MANAGEMENT STRATEGIES IN NGEO SATELLITE NETWORKS

Taking the features of NGEO satellite networks into consideration, various routing and mobility management techniques are proposed for satellite networks. The main ideas behind these proposals can be classified as follows:

1. To handle dynamic topology changes with minimum overhead. For this purpose, mainly periodicity and predictability of the constellation topology are considered. Frequent handovers and limited storage and processing capabilities of satellites are challenges to cope with.

2. To prevent an outgoing call from dropping due to link handover as the satellite topology changes. For this purpose, some proposed routing algorithms aim at reducing the probability of link handover occurrence. Both periodicity and predictability of topology changes are mainly considered.

3. To minimize the length of the paths in terms of propagation delay and/or number of satellite hops, in order to avoid poor resource utilization as well as high end-to-end delay. Constant size, highly symmetric and uniform nature of constellations constructs some advantageous features that should be considered.

4. To prevent congestion of some ISLs, while others are idle. For this purpose various load balancing algorithms are proposed. As a common feature, these algorithms are adaptive to dynamic and non-homogeneous traffic. Load balancing algorithms could also benefit from symmetric and uniform characteristics of constellation topologies, and must consider physical restrictions of satellites.

5. To perform traffic-based routing in order to satisfy quality of service (QoS) requirements. In this context, main problems are dynamic topology and traffic, frequent handovers, and physical limitations of satellites.

6. To provide better integration of satellite networks with terrestrial networks. In this context, some works try to apply existing routing algorithms to satellite constellations in order to provide easier integration of the satellite network to the terrestrial network. As an example, some works examine how to adapt IP routing to satellites, so that the constellation can be seamlessly integrated into the Internet. On the other hand, if the satellite network uses its own arbitrary routing protocol, the problem to solve is how to integrate it with the terrestrial networks.

7. To perform efficient multicasting over satellites, regarding the characteristics of satellites in broadcasting.

Table 3.1. Relationship between classes of routing objectives and satellite features

Effect	**Features Particular to Satellite Systems**	**Aim of Routing Techniques**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Satellite Mobility	dynamic network topology	✓	✓	✓		✓	✓	✓
	predictable & periodic topology changes	✓	✓	✓			✓	✓
	dynamic & nonhomogeneous traffic				✓	✓		✓
	handovers	✓	✓			✓	✓	✓
Limitations & Capabilities of Satellites	limited power and on-board processing capacity	✓			✓	✓	✓	✓
	difficulty of implementing latest technology						✓	✓
	broadcast nature							✓
Nature of Satellite Constellation	higher propagation delays			✓				✓
	fixed number of nodes			✓	✓		✓	✓
	highly symmetric and uniform structure			✓	✓			✓

To achieve each of these seven objectives, some of the aforementioned features of satellite systems should be considered. Table 3.1 shows the relationship between classes of

routing objectives and satellite features. The check marks indicate the existence of an effective relation but are not necessarily restricted to those shown in Table 3.1.

Now, we are going to examine how the proposed routing and mobility management techniques may achieve the objectives stated above, discuss some of their advantages and disadvantages, and point out open research areas in this context.

3.1. Handling Dynamic Topology

Due to the rapid changes in the status of the links and satellite positions, a satellite network can be considered as a dynamic-topology network. The utilization of conventional routing techniques widely used in terrestrial networks (such as Open Shortest Path First (OSPF) [19] and Routing Information Protocol (RIP) [20]) is not feasible for satellite networks since these protocols rely on the exchange of topology information that must be constantly refreshed, which incurs substantial overhead. However, although the topology of a satellite network rapidly changes, these changes are periodic and predictable because of the strict orbital movements of the satellites. Therefore, some routing techniques are proposed utilizing this periodicity feature. In [21], a LEO satellite network is modeled as a finite state automaton (FSA), where the system period (which is the least common multiple of the satellite layer's orbital period and the Earth period) is divided into states. The states are derived from the ISL connectivity data, so that the network has a fixed topology in each state. Due to the periodicity of the constellation topology, there can only be a finite number of states. Then, it is proposed to perform optimal routing on each of these fixed topologies for the best use of ISLs in the system**.** Werner *et al.* [22] proposes Dynamic Virtual Topology Routing (DVTR) for ATM-based satellite networks, which work in similar way as the FSA algorithm. Again, system period is divided into a set of time intervals, so that topology remains constant over each interval. For each interval, the best path can be found by totally off-line optimization procedure, or selected from a set of alternative paths depending on the on-line traffic information. In these FSA-based techniques, a number of routing tables are stored on-board and retrieved when topology changes. Although the messaging overhead and computational complexity is reduced, large storage capacity is needed in the satellites. In order to reduce the on-board storage

requirements, a suitable number of network control centers (NCC), which are located on the ground, can be used [23]. In this context, deciding on the number of NCCs to use and their distribution on the globe are open issues.

Another concept worth mentioning, which is tailored to dynamic satellite constellation is the virtual node (VN) concept [24]. A logical address is assigned to the fixed portions of the Earth's surface. Then, by using *Earth-fixed* satellite systems described in Section 1.1, a satellite embodies the VN above this fixed Earth footprint for the time period during which it is serving that footprint. Each VN is embodied at any given time by a certain physical satellite. As a satellite disappears over the horizon, its corresponding VN becomes represented by the next satellite passing overhead and the state information (such as routing table entries or channel allocation information) is transferred to it. Handovers between VNs and physical satellites are synchronously performed; hence, the virtual topology remains unchanged. A routing decision is made on this fixed virtual topology, and consequently, the network layer is isolated from the satellite constellation dynamics.

Recently, many routing protocols are proposed based on VN concept. A mechanism to adopt IP routing at the VNs in order to seamlessly integrate space network with terrestrial Internet and provide direct support for IP-QoS and IP-Multicast is presented in [25]. A distributed datagram routing algorithm, and a multicast routing algorithm regarding satellite network as a mesh topology consisting of fixed logical locations (virtual nodes) are introduced in [26], and [27], respectively. Moreover, some hierarchical routing algorithms that are developed for integrated satellite networks consisting of LEOs and MEOs [28, 29] or HAPs, LEOs and GEOs [30] simplify LEO layer by modeling it as a fixed virtual network. Although VN concept is widely accepted, it has some deficiencies, which come from the fact that VN concept necessitates one-to-one correspondence between physical satellites and virtual nodes and it could not be applied for systems where more than one satellite can serve for a single footprint area. We describe major shortcomings of VN concept in Section 5.2, and to make up with these, we propose and model a Multi-state Virtual Network (MSVN) topology which enables more than one satellite to cover a single footprint area. MSVN is described in detail in Section 5.3 together with its potential advantages.

Using VN technique, network layer handover could be totally eliminated because VNs have fixed network layer addresses and routing mechanisms are utilized over the fixed virtual topology. However, although there is no need for network layer handover, physical network is dynamic and handover in lower layers could result in significant packet loss. Designing a smooth (low packet loss) and fast (low latency) handover algorithm is a crucial issue. Especially for satellite environments with long propagation delays, system performance could be significantly improved by using proper handover mechanisms. Therefore, we propose an efficient link-layer handover scheme for VN-based satellite systems, namely Virtual Node Handover (VN-HO) algorithm, in Section 5.4. Next, we propose soft handover algorithm and semi-soft handover algorithm for MSVN-based satellite networks; MSVN-SHO and MSVN-SSHO. Comparison of proposed algorithms shows possible advantage of MSVN over single state conventional VN architecture.

3.2. Reducing Link Handovers and Rerouting Issues

The issue of (re)connection setup overhead when a connection is broken is imperative for satellite networks due to the highly dynamic nature of the network topology. When some ISLs are switched off or the up/down link (UDL) between the terrestrial node and the corresponding initial satellite is broken, handover is needed to maintain the active connections. Rerouting attempts during link handovers cause delay jitter and signaling overhead. Moreover, because of the possibility of resource unavailability in alternate paths and the delay caused by rerouting, the forced termination probability of ongoing connections is increased. Therefore, it is desirable to minimize the possibility of rerouting due to handovers. It is especially important for real-time multimedia applications to maintain QoS guarantees using a connection-oriented routing protocol. Thus, minimizing number of connection handovers is a crucial issue.

In order to reduce link-handovers, Werner *et al.* [31] propose an optimization procedure for their system based on DVTR mentioned in previous section. While calculating the most appropriate path between a satellite pair for each time interval, minimizing the number of hand-overs in the whole system period is taken into account,

while keeping delay and/or delay jitter minimal. Optimization procedure results in a unique route for all connections between a satellite pair during a time interval. The algorithm in [31] is improved in [22] such that the optimization is done over a sliding time window, rather than the whole system period. In other words, the routes are determined such that hand-over rate and hand-over delay jitter occurring in a time window is minimized. By sliding the window, new routes are determined after each topology change. Relative magnitude of call duration to the window size has an important effect on the performance of the algorithm. It is stated that the window size is fixed and it should be around the average call duration to achieve better performance.

Ercetin *et al.* [32] propose a predictive routing protocol which aims to reduce handover probability in an online fashion while taking traffic characteristics into account. In this approach, traffic load on the ISLs up to a short time in a future is predicted using the deterministic nature of the satellite topology and user location information. Then, for each connection, k ordered paths are obtained depending on the residual bandwidth information. This operation is done for l short intervals up to a time that the footprint of a satellite completely changes. Then an appropriate path is selected among k paths for each interval such that link changes are reduced (hence overhead due to handovers is decreased) as well as the user traffic is balanced. Computational overhead of this algorithm is quite high, especially for large number of k values.

Jukan *et al.* propose to reduce the handover ratio by favoring ISLs with higher lifetime [33]. For each connection request, request packets are flooded towards the destination. While traversing the route, these packets gather the lifetime information of intermediate satellites, and this information is used by the destination node for deciding on the most appropriate path. Chen *et al.* [34] also consider minimizing handovers in their proposed adaptive routing scheme. Among the set of paths that satisfy the QoS requirements, a path which can minimize the possible number of handovers and also which is not poorer than the best possible path with a predefined degree is selected.

The mentioned procedures make optimization between two satellite nodes. Hence, only ISL handovers are considered. However, a connection should also be reestablished

when the UDL between the source ground station and the ingress satellite, or between the destination ground station and the egress satellite is broken. This kind of handover is defined as *intersatellite handover* [35] and should be considered for better optimization. It is preferable to reduce intersatellite handovers due to the movement of satellites with respect to user stations. [35] introduces a probabilistic routing protocol (PRP), which tries to reduce rerouting between two ground end-users by utilizing LEO satellite topology dynamics and call statistics. Basically, the algorithm tries not to use ISLs that would be switched off before the connection is over. Since exact call duration is not known a priori, the probability density function (pdf) of the time duration in which the call uses the established route is utilized. The determined pdf is used to establish the routes of the connection, such that the routes are terminated by a call termination event or an intersatellite handover instead of a link handover with a target probability p. A distinction is made between route calculation for newly arrived calls and intersatellite handover calls since dropping an ongoing call that experiences an intersatellite handover is more annoying than blocking a new call. Since PRP increases call dropping rate, using PRP is suggested only for calculating the routes for newly arriving calls. For intersatellite handover calls, Footprint Handover Rerouting Protocol (FHRP) [36] is used. FHRP balances the simplicity of route augmentation and optimality of complete rerouting. It has two phases: Augmentation phase and Footprint Rerouting (FR) phase. When either source or destination satellite goes out of visibility region of ground terminals, a route between new end satellite and the original route is established, and the unused portion is removed. This is called augmentation. If it is not possible to do so, the connection is rerouted using the original routing algorithm. FR is only possible when new end satellites are the successors of the end satellites in the original route. When both ends experience handover, FR is applied and the connection route changes completely.

3.3. Path Minimization Algorithms

In a satellite network, the cost of a path is determined by the propagation and processing delays on the satellites. When compared to terrestrial networks, the propagation delay is more important in space networks due to height of the satellites and long distances between the nodes. Moreover, as more number of satellite hops is traversed, total processing and propagation delays increase. Reducing the processing delay has some

beneficial side effects like reducing the data blocking probability and yielding better power consumption in satellites. Therefore, path minimization is a crucial task in satellite networks.

In some works, it is assumed that ISLs have fixed lengths. Various authors argue that this assumption is reasonable because in most of the constellation topologies, the length variation is low, especially for crowded regions which are near to the equatorial region. Moreover, it is also claimed that minimizing the hop count is more critical for improving the performance of the system and therefore it is reasonable to ignore dynamic length variations of ISLs. On the other hand, numerous authors think that the dominating factor for performance is the propagation delay and they aim to find a minimum-propagation-delay path with the minimal hop count among the paths. This complicates the task, since ISL lengths change with time due to the dynamism of the topology. The predictability of topology changes and the known facts about the nature of satellite networks (e.g. links over the polar regions are shorter than links over the equatorial region) can be used to simplify this task. Nevertheless, extra storage and processing complexity is required to consider the propagation delays.

Sun *et al.* [37] deal with static routing in a regular LEO satellite network, which is modeled as two dimensional N-ary hypercube. The minimum-hop path is found by Dijkstra's algorithm and some contention resolution schemes are investigated for maximizing the throughput. It is shown analytically and validated by simulation results that a scheduling scheme favoring packets closest to their destinations results in maximum system throughput, at the expense of degraded system fairness. Actually, there could be many minimum hop paths in a satellite constellation, due to its symmetric and uniform nature. Therefore, it is proposed in some works to favor the one with the minimum propagation delay. The most trivial way to do this is to store the length information of all links for a system period in each satellite (or in ground stations that perform routing) and to apply a shortest path algorithm using this information. This necessitates a high storage capacity. Henderson *et al.* [38] and Ekici *et al.* [26] develop distributed algorithms for minimizing the propagation delay. The geography-based algorithm of Henderson *et al.* is based on the hypothesis that the series of locally optimal forwarding decisions will yield a route that is close to the optimal route. In other words, depending on the geographic

information embedded in the addresses, each satellite forwards the packet to its neighbor that most reduces the distance to the destination. On the other hand, Ekici *et al.* [26] introduce the datagram routing algorithm (DRA) for an idealized polar constellation. It regards the satellite network as a mesh topology consisting of logical locations (virtual nodes). Data packets are routed in a distributed fashion in this fixed topology. DRA consists of two phases: At a given satellite hop, first it finds all the neighboring satellites that can move the packet one hop closer to the destination. Then, from the candidate next hops, it selects the one which most reduces the remaining distance to the destination.

There are many other algorithms that utilize information on expected traffic characteristics and handover possibilities of ISLs while applying a shortest path algorithm or similarly consider dynamic traffic characteristics while deciding on the most appropriate path among a set of shortest paths. The objective of the algorithms from the latter group is mainly distributing the traffic load in a more balanced way. They are described in the next section.

3.4. Load Balancing Algorithms

Since population distribution on the Earth surface is highly non-uniform, traffic requirements are unbalanced in a low orbit satellite network. This may lead to congestion in some resources, while others are under-utilized. To overcome this problem, the routing algorithm should distribute the flows in a balanced way over appropriate ISLs between the communicating nodes.

We classify these algorithms according to the place where the routing is performed: Source-based, central, distributed and hierarchical load balancing algorithms, as shown in Table 3.2.

3.4.1. Source-based Load Balancing

In source-based load balancing algorithms, the route to a given destination node is calculated by the ingress node. The ingress nodes could be a terrestrial node or a satellite.

If it is located on the ground, an extra signaling delay is introduced. However, in the latter case, computational and storage requirements to perform route calculation can exceed the capacity limits of a satellite.

Frank *et al.* [39] classify source-based load balancing algorithms further as isolated and non-isolated algorithms. Isolated algorithms use only information local to the node where routing is performed. In non-isolated algorithms, traffic information is gathered from the whole network. Authors suggest a non-isolated algorithm as follows: Each node keeps the graph of the whole network. When routing is performed, all nodes and edges that are near to saturation point are pruned. Then a shortest path algorithm is run considering the propagation delay as well as the constant transit delay per hop.

Chen *et al.* [34] propose an alternative adaptive routing algorithm that uses the information on both the average and the minimum number of occupied channels per route. First, the algorithm finds a set of candidate minimum delay paths that also minimize the handover probability and delay jitter. Then, among these alternate paths, the one with minimum traffic weight, which is determined by a weighted combination of average and minimum number of occupied channels over the route, is selected.

Table 3.2. Comparison of different load balancing schemes

LOAD BALANCING SCHEME		ADVANTAGES	DISADVANTAGES
SOURCE-BASED	ISOLATED	● Simple	● No global view of the network ● Low utilization
	NON-ISOLATED	● Global view of the network ● Good traffic adaptiveness	● Difficult to guarantee up-to-dateness of the traffic information ● High signaling overhead
CENTRAL		● Global view of the network ● Whole information can be used for an overall optimization procedure. ● Computational complexity is carried from satellites to a central node	● Difficult to guarantee up-to-dateness of the traffic information ● High signaling overhead ● Scalability problem
DISTRIBUTED		● Each node uses up-to-date local information ● Low signaling overhead ● No rerouting issues ● Fast adaptation to traffic changes	● No global view of the traffic load distribution ● Utilization is somewhat low
HIERARCHICAL		● More routing choices ● Better adaptation to traffic changes with less computational and signaling cost.	● Physical challenges in providing inter-orbital satellite communications ● Increased system complexity

Non-isolated routing technique increases the computational and signaling complexity of the routing architecture, but it is superior to isolated algorithms since it considers traffic adaptivity in the whole network. However, there is a potential drawback of non-isolated algorithms: The gathered traffic information may not reflect the actual condition since (a) the information takes time to be distributed in the constellation (due to high propagation delays) and (b) in order to avoid excessive signaling, state changes in the network are not always advertised.

3.2.2. Central Load Balancing

In central load balancing algorithms, routing tables are calculated in a central node and then stored in the satellite nodes. This central node can be a satellite or terrestrial node. Mainly, we can consider *optimal routing* algorithms in this context. Optimal routing algorithms are network-oriented, aiming at minimizing the mean blocking probability in the network by providing better load balancing.

The aforementioned FSA-based algorithm of Chang *et al.* [21], which is offline, assigns expected traffic loads to links depending on the statistical information on the potential requirement density for each node and the distance between the nodes. For each state, it aims to maximize the minimum residual capacity in the network. Since this is an NP-complete mixed integer optimization problem, it uses some heuristics to provide optimal routing. Authors compare this optimal routing algorithm with a dynamic routing approach, which is based on shortest path algorithm. In optimal routing algorithm, routing tables are updated as states change, whereas dynamic routing updates routes in every broadcasting period, using the obtained link status information. Authors conclude through simulation that optimal routing is superior in terms of newly initiated and ongoing call blocking probability.

Papapetreu *et al.* [40] propose to perform the flow deviation (FD) algorithm [41], which is a well known optimal routing algorithm that aims to find a routing pattern that minimizes the mean network delay. Depending on the information gathered from the whole network, a designated central node performs the FD algorithm. The FD algorithm

splits the load to different paths according to path lengths, which are defined as a flow dependent metric. As the lengths of the paths change, the FD algorithm continuously adapts these changes by deviating traffic from one path to another, so that the defined cost metric is minimized. The authors show via simulation that the FD algorithm is superior to Dijkstra for finding the path minimizing the propagation delay and Adaptive Dijkstra for finding the path minimizing a flow metric.

By performing routing in a central node, better traffic engineering could be maintained using the global view of the network. However, central load balancing algorithms share the same problems with non-isolated source-based routing algorithms. Computational complexity can be carried to a ground node that does not suffer much from power limits, but the high signaling requirement and the difficulty of accurately transferring traffic information are the most challenging problems. Moreover, the centralized routing approach may present scalability problems due to the capacity limits of the central node, and the rapid increase in the computational complexity with the enlargement of the network size.

3.2.3. Distributed Load Balancing

Because of the highly symmetric and uniform structure of satellite constellations, there can be many minimum-hop paths between two satellites and routing can probably be done efficiently. Establishing a static connection between two nodes may lead to poor utilization of alternate paths. Moreover, as we mentioned before, the connection-oriented approach may suffer in attaining path connectivity by handover mechanisms. Rather, a distributed next hop routing strategy seems to be simpler. Each satellite independently decides on the best next hop to forward the packet. Ekici *et al.* [26] implement this approach in the aforementioned datagram routing algorithm. The main objective of the algorithm is minimizing the propagation delay, but a satellite may change its decision if the output queue for the ISL over the minimum propagation delay path is congested. Taleb *et al.* [42] claim that a better load balancing might be achieved, given that a satellite is aware of the traffic conditions at the next hop satellite. They propose an explicit load balancing (ELB) scheme, where a congested satellite sends a signal to its neighboring satellites to

decrease their sending rates, and its neighbors search for alternate paths. This method reduces the packet dropping probability but it is not safe from signaling congestion due to feedback packets (even though signaling packets are sent only when it is necessary, they could be needed very frequently in some conditions). Algorithms in [26] and [42] do not take any action for load balancing until some nodes experience a certain level of congestion, i.e. they are not proactive. We support the idea that it is more appropriate to avoid congestion before it happens and we provide a priority-based adaptive routing algorithm (PAR) in Chapter 4, which aims to balance the traffic before any congestion occurs. Since there may be many minimum hop paths between a source-destination pair in a satellite network, for each intermediate satellite, there may be more than one outgoing links that are on one of these minimum hop paths. When a satellite node receives a packet, among these ISLs, it selects the one with the highest priority. If the ISL with the highest priority is congested at that instant, then the ISLs with lower priorities are selected. If all of the ISLs (that are on a minimum hop path) are congested, then packet is dropped (or deflected [13]). Priorities of links dynamically change depending on the past utilization and queuing information. We compare PAR with other adaptive minimum hop path algorithms fixed adaptive routing (FAR) and random adaptive routing (RAR). At each hop, among the ISLs that are over a minimum hop path, FAR first selects the one that is towards a given direction (vertical or horizontal). If that link is congested, the other direction is selected. RAR makes the selection of initial ISL candidate in a random manner. Simulation results show that using PAR algorithm not only increase throughput, but also decrease queuing delay. Moreover, PAR algorithm does not need any signaling overhead due to feedback packets.

The distributed load balancing algorithms mentioned above provide fast reaction to traffic changes when compared with the source-based and centralized load balancing algorithms. However, they use only the local traffic information, which might not reflect the entire traffic load distribution. Surely, it is possible to distribute the local information to the whole network and use it in local next-hop decisions, but this will cause extensive signaling overhead.

Another approach that could be considered in the context of distributed load balancing is applying Ant Colony Optimization (ACO) algorithm in LEO satellite

networks [43]. In ACO algorithm, simple agents (called ants) are emitted by satellite nodes with a given period. These ants gather delay information along paths through the system and store it within the routing nodes on their return to their source node. Nodes on good paths will be visited frequently by ants reporting small trip times, thus reinforcing routing table entries for links contained in those paths, and diminishing those of the other links. Ants on poorer paths will arrive later and report larger delays, causing routing table entries for such links largely unchanged. In ACO algorithm, there is a trade-off between emitting frequency of ants, and signaling overhead. Emitting ants more frequently yield better adaptation to traffic changes, but may congest the traffic. Ant-based algorithms are especially appealing for ad-hoc space networks that are out of scope of this thesis.

3.2.4. Hierarchical Load Balancing

Hierarchical (multilayered) satellite architectures with inter-orbital links (IOLs) between layers of satellite constellations are of much interest as they may yield better performance than individual layers. Figure 3.1 depicts a layered satellite architecture with some applications. The hierarchical approach aims to reduce the computational complexity on the satellites and the communication load on the network when compared with non-isolated algorithms, while enabling better adaptation to traffic changes.

Lee *et al.* [44] propose a satellite-over-satellite (SoS) system, where the satellite architecture is composed of LEO and MEO layers. Lower layer (LEO) satellites send ISL state messages to upper layer. An upper layer (MEO) satellite uses this state information in order to derive some local routing information (LRI) about the LEO satellites which are in its coverage area (these change with time due to the relative mobility of LEO satellites with respect to MEO satellites). This information is also exchanged between MEO satellites. In addition, MEO satellites derive global routing information (GRI) including all routing information of the LEO and MEO layer satellites by using this exchanged state information and send it through IOL to all LEO layer satellites that are within their coverage area. In the proposed routing algorithm, short-distance-dependent traffic is transmitted through lower layer satellites, but long-distance-dependent traffic is transmitted

through IOL up to the MEO layer in order to minimize the average number of satellite hops and resource consumption.

The satellite grouping and routing protocol (SGRP) proposed in [45] is another hierarchical algorithm where LEO satellites are divided into groups according to the footprint area of the MEO satellites in each snapshot period. Each LEO group is managed by a MEO. Similar to the proposal in [44], a MEO satellite computes the minimum-delay paths for its LEO members, depending on the link state information arriving from the LEOs. The authors provide a detailed analysis of their proposed system.

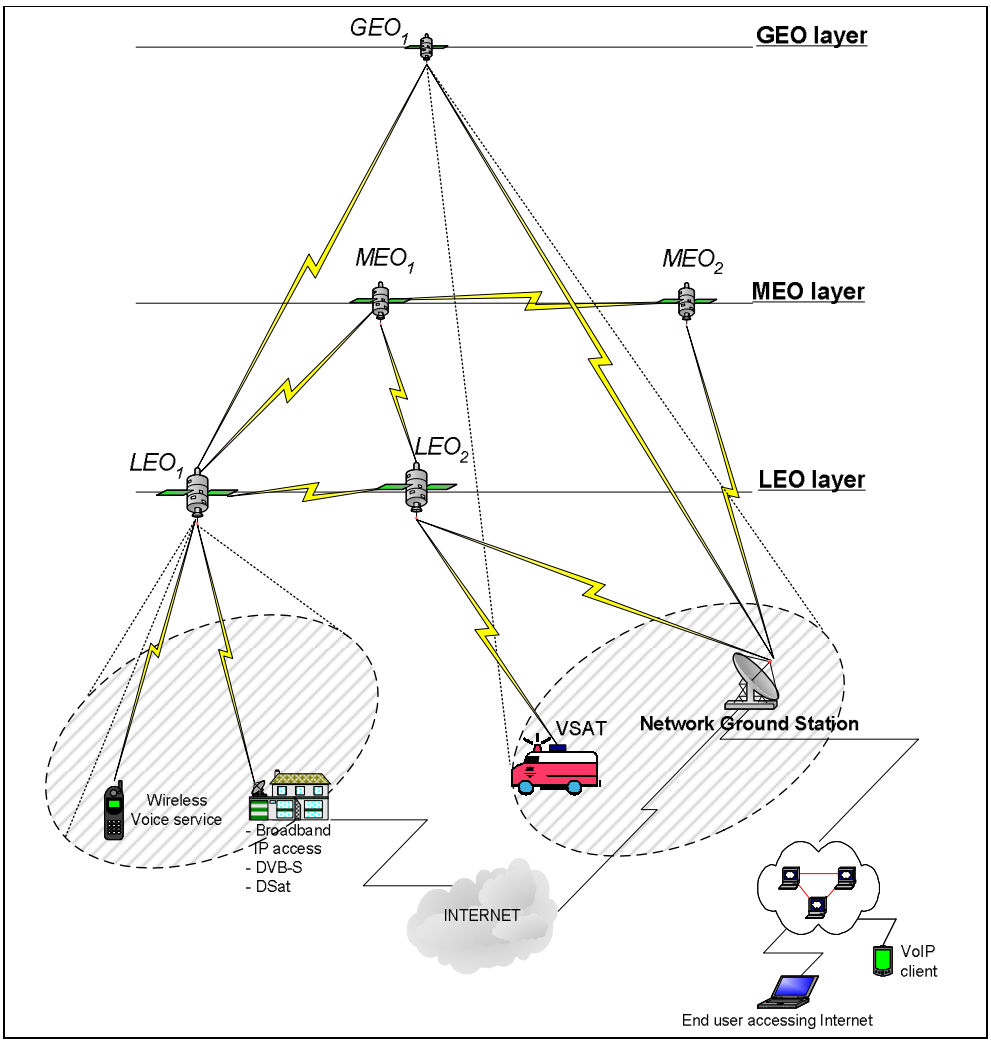


Figure 3.1. A layered satellite network with sample applications

Dash *et al.* [30] consider a three-layered architecture consisting of GEOs, LEOs and HAPs for Voice over IP (VoIP) application. GEOs act as backbone routers, LEOs as the second layer and HAPs to cover special areas with high and sensitive traffic such as battlefields and disaster areas. LEO layer satellites are assumed to be Earth-fixed and modelled as virtual nodes; hence GEO satellites cover logically fixed LEO topologies. LEO satellites exchange their routing tables as their footprint area changes. This architecture enables all of the satellites to see other layers stationary. LEO satellites and HAPs measure residual bandwidth on their outgoing links and send the information to GEO layer with a given period. Since GEO satellites have limited onboard processing capacities, this link state information can be onto the fixed terrestrial gateways for processing. After forming intra-domain routing tables, gateways upload these tables to GEO layer, and GEO satellites flood them to the LEO satellites and HAPs that are in their coverage area. Also an aggregated routing table for each domain is formed which includes the maximum residual bandwidth paths between different border nodes of the domain. These tables are then exchanged between GEOs and transferred to the lower layers.

NGEO satellite networks integrated with HAPs may have great potential in the next generation telecommunication services. In [30], mobility of the LEO layer is simplified by assuming that the physical LEO satellite network can be reduced to a fixed logical topology. As the authors indicate, this assumption is valid for the satellite systems with Earth-fixed footprints. However this is impractical in most satellite systems, and mobility of the satellites should be handled in more realistic ways. In Chapter 6, we consider internetworking between HAPs and NGEO satellites without discarding the mobility of satellites. We consider an integrated architecture, where HAPs and satellites communicate via high capacity free-space optical links. We focus on establishment of optical links between HAPs and mobile satellites with limited resources, in order to maximize availability and performance of the system. We formulate an optimization problem for matching HAPs and satellites in such a way that the utilization of HAPs is maximized together with the average elevation angle between HAPs and satellites. We also propose a method to avoid frequent switching of optical links.

The advent of hierarchical architectures implies more redundancy and routing choices in satellite systems. A variety of topological design and routing issues should be

investigated for enabling the best usage of satellite technologies in the future communication systems.

3.5. Traffic-based Routing

In order to support the rapidly-growing real-time multimedia services, satellite systems should be able to offer QoS guarantees, which is difficult in connectionless networks, in particular due to the difficulty in accounting for the delay aspects of QoS and sequencing. Usually, QoS guarantees are provided through connection-orientation. However, due to the high mobility of satellites in NGEO systems, it is difficult to attain path connectivity. Therefore, some of the algorithms described above aim at minimizing the rerouting probability due to handovers while calculating the routes. Nevertheless, it is not easy to offer QoS guarantees without reducing the rerouting probability to very low levels. The VN concept can be used in order to get rid of topology changes. However, this approach also has its drawbacks as described before. Therefore, this topic needs further study.

Another issue in providing QoS guarantees is to reduce delay jitter that occurs due to path rerouting. In a LEO satellite network, the movement of satellites causes changes in relative positions between any two satellites that belong to different orbits. This may result in unacceptable levels of delay jitter. Therefore, a routing algorithm has to try to reduce the delay jitter for better QoS conditions while keeping the delay itself as low as possible at the same time. This issue is considered in the context of various proposed algorithms in the literature [22, 34]. Regarding handovers, it is expected to choose a new path that is not much different from the previous one in its length, although sometimes the selected path is not the best (shortest) one.

Jukan *et al.* [33] propose a distributed QoS-routing approach. The source node floods connection request packets towards the destination. At each intermediate node, the quality parameters (delay, lifetime of ISLs etc.) are updated. When connection request packets reach the destination node, the destination eliminates the paths which do not satisfy the

QoS requirements. Among all feasible paths, the one with the minimum number of hops and maximal lifetime is selected.

Kandus *et al.* [46] propose a traffic-class-dependent (TCD) routing algorithm. Three classes of traffic are considered: A: delay-sensitive traffic, B: throughput-sensitive traffic, and C: best-effort traffic. The routing algorithm behaves differently for each class of traffic. Each satellite has three separate outgoing queues (one for each traffic class) serving for each outgoing link. A scheduler should be implemented which defines the actual transmission sequence of packets in outgoing queues. Obviously, the selection of the scheduling policy has a big impact on the routing performance of the particular traffic class. Therefore, the authors investigate five different scheduling policies, which we do not describe here. The TCD algorithm attempts to guarantee QoS for different traffic classes. However, it may assign a single route for a specific class with huge data traffic and may heavily overload the chosen path. This may negatively affect the traffic load balancing over the entire constellation.

3.6. Routing from Space-Ground Integration Point-of-view

The problem of how to integrate space networks with terrestrial networks arises for the purpose of using satellite systems as a part of a global communication system (such as Internet). Basically, there are two trends in this context: according to the first trend, the goal is to apply the existing algorithms as extensively as possible and provide interfaces with terrestrial networks as easily as possible. Conversely, there is a second trend to use an arbitrary routing protocol in the space segment. In other words, the satellite network can be designed and operated independently of the terrestrial network. The disadvantage of the latter trend is that an address resolution protocol and some complex interworking functions are needed. However, it is still a better approach than the former trend in terms of scalability.

Currently, IP protocols dominate the end systems attached to the satellite terminals. Therefore, research that investigates how to apply IP routing directly to satellite systems can be considered in the context of the first trend. Wood *et al.* [25] examine a strategy

(based on the VN concept) that aims to adopt IP routing at the satellites. This strategy permits direct support for IP-multicast and IP-QoS (integrated and differentiated service models). However, there are some challenging problems with this technique. First, the variable-length IP packets should be fit into a fixed-length frame structure in the space segment. The authors propose to achieve this by using either explicit IP-level fragmentation or implicit lower-level fragmentation in order to break packets, such that they can fill in the frames, and by using padding in order to fill up the frame structures. The second important problem is scalability. When the terrestrial network increases in size, a large amount of routing information must be updated for both the terrestrial and the satellite networks. However, this is not feasible for satellites, since the capacity of satellites is limited and cannot be upgraded once they are launched. Therefore, it is better to separate and isolate satellite and Internet routing updates. Finally, since IP-routing is slow and needs high processing power, it is not suitable for satellites which are equipped with limited on-board processing capacities. The authors argue that the IP-routing performance is continuously improving and by using some shortcut IP-switching techniques, such as multi-protocol label switching (MPLS), it seems possible to overcome this problem. We briefly discuss MPLS over satellite constellations in the last part of this section.

While IP protocols are dominant in terrestrial nodes, majority of the proposed satellite systems (e.g. Spaceway, Astrolink, Skyway, and Cyberstar) plan to use ATM as the link layer technology for interconnecting the satellite terminals [47]. This is partly due to the fact that, at the time of designing these systems, ATM was seen as the dominant future network technology. There is a significant time gap between the design and the operational stages of a satellite system. That lag usually makes it difficult to have and to utilize the state-of-the-art technology in orbit. Moreover, next generation satellite networks are expected to provide and support multiple types of services and to interwork with different terrestrial networks, such as B-ISDN, Internet, etc. Therefore, it is reasonable to isolate routing in the space segment from the terrestrial networks. For this purpose, it is suggested to view the satellite network as an autonomous system (AS) with a different addressing scheme. In order to integrate satellite constellation in the Internet, at a border gateway (BG), IP address of the exit node is translated to a network address via an address resolution protocol.

An external routing protocol (such as Border Gateway Protocol (BGP)) could be run on the BG of this AS, in order to communicate with terrestrial ASs. To reduce the load on the satellites, it is more advantageous to implement BGs in a dedicated ground station. In terrestrial networks, the cost metrics between different ASs are the dominating factor while establishing a route between nodes from different ASs. This is because generally the internal paths in a terrestrial AS are quite shorter than the inter-AS paths. However, this is not the case for a satellite network, since an internal path can be easily as long as an external path. Thus, in the satellite context, the routing protocol should consider the internal cost metric to be as important as the external one.

Wood *et al.* [25] claim that in the future, satellite constellations can carry IP traffic by using a combination of border routing protocols, tunneling, network address translation and MPLS. MPLS allows adopting new paradigms for conventional IP traffic by decoupling packet forwarding from the information carried in the IP header. This is achieved by distributing the routing information to the core routers and assigning short labels to the packets at the ingress point. MPLS has some appealing mechanisms essentially supporting the integration of the IP world with QoS and traffic engineering features. Donner *et al.* [48] deal with how to adopt MPLS over a satellite constellation. Constellation topologies that do not have any seam (inclined Walker constellations) are seen as the strongest candidates to host MPLS, due to their permanent nature. Nevertheless, the inherent and frequent handovers between ground and satellite stations and topology dynamics due to varying ISL lengths remain a challenge. At the ingress point of an MPLS network, there are label edge routers (LER) which manage the label distribution and, in some cases, perform route computations. The authors propose to locate LERs on the ground in order to avoid i) expensive and complex on-board processing in the satellites, and ii) extra signaling overhead needed for restarting QoS negotiation or admission control for rerouting of a label switching path (LSP). Different scenarios for route computation (including “intelligence” in terms of traffic engineering, adaptiveness, and/or optimization) and LSP management (establishing the result of “route computation” in the network) are investigated. Centralized routing approaches are viewed as more promising for use within an MPLS framework. The interested reader may refer to [48] for further information. MPLS-based networking in satellite constellations is an appealing

approach. However, some important practical problems related to rerouting and maintenance overhead are still unsolved and deserve further study.

3.7. Multicast Routing

Given the ability of satellites to broadcast large amount of data over a very large area, multicasting over satellites have become very hot research topic. Ekici *et al.* [27] indicate, none of the existing multicast routing protocols (like reverse-path multicast (RVM) [49], distance vector multicast routing protocol (DVMRP) [50] or multicast routing extensions for OSPF (MOSPF) [51]) are suitable for satellite networks, since they employ some type of periodic message exchanges to form or maintain multicast trees, and this is not favorable due to the physical limitations of satellites. To fill the gap, authors develop a multicast routing protocol for LEO satellite IP networks, where multicast trees are formed based on datagram routing algorithm [26]. The algorithm aims to minimize number of branches going out of a satellite at each step. Authors conclude that their algorithm outperforms existing multicast routing algorithms in terms of end-to-end delay. But, multicast routing algorithms for multilayered satellite networks is still an appealing research area.

3.8. Summary

Satellite communication systems have some intrinsic features that significantly affect the performance of their routing algorithms. Particularly, adopting traffic and topology dynamics may incur significant overhead in the course of utilizing satellite resources. Various routing algorithms have been proposed for satellite networks in order to overcome this drawback. In this chapter, we classified these algorithms and described relevant technical issues for use in the next generation satellite networks. Although some algorithms may seem to meet performance criteria in certain cases, routing algorithms that supports minimum overhead with better resource utilization remains a practical problem for the next generation. Both QoS and multicast routing algorithms are still formidable tasks in satellite networks. Furthermore, the advent of sophisticated satellite network

architectures such as multilayered systems and new space technologies such as HAPs continue to broaden potential routing choices.

4. PRIORITY-BASED ADAPTIVE ROUTING IN NGEO SATELLITE NETWORKS

4.1. Motivation and Related Work

With the rapid globalization of telecommunications industry, satellites are expected to widely appear in future telecommunication systems, due to their capabilities such as extensive geographic coverage, inherent multicast capabilities and support for mission-critical applications. GEO satellites suffer from high propagation delay, which is not suitable for most applications (especially for real-time applications). Therefore focus has been directed towards development of non-geostationary (NGEO) systems consisting of LEO and MEO satellites. For better utilization of satellites and increasing the performance of the system, new NGEO systems usually support ISLs between satellites. The use of ISLs raises the issue of routing in the satellite network. Employing an efficient routing protocol is a critical issue, since satellite network resources are costly and must be managed in an optimal way. Unfortunately, mobility of satellites complicates the routing issue in an NGEO satellite system; hence routing algorithms used in terrestrial networks are not directly applicable in satellite networks. Satellite movements cause both the dynamicity of network topology (variation in ISL lengths, etc.) and dynamicity of traffic passing over satellites. A good routing algorithm should be adaptive to these dynamics.

It is important to note that changes in the network topology are periodic and predictable because of the strict orbital movements of satellites. Therefore it is reasonable to use this periodicity feature while calculating routes. Dynamic Virtual Topology Routing (DVTR) [22] is one of the most common routing methods that use the periodicity of the topology changes. DVTR divides the system period into a set of time intervals, so that topology remains constant over each interval. Time intervals are chosen to be short enough to assume that costs of ISLs are fixed. Therefore optimal paths and alternate paths can be established using well-known methods like Dijkstra's shortest path algorithm. DVTR approach decreases the online computational complexity with the expense of large storage requirements.

In a satellite network, routing decisions can be made offline or on-board. In the former case, the routing algorithm could use information about predictable topology changes (such as changes in the ISL lengths and connectivities), but it would not be adaptive to traffic load changes. However, on-board routing algorithms yield better efficiency for dynamic traffic cases with the expense of increased complexity. Several adaptive routing protocols that take the traffic characteristics into account have been proposed for NGEO constellations. In [52], satellite constellation is modeled as a regular mesh and ISLs are assumed to have fixed length. The work deals with adaptive routing with a limited set of alternative routes. However, there may be many shortest paths (in terms of hop-count) in a mesh-like network which can be fully utilized. In [26], Ekici et al. propose the Datagram Routing Protocol, where ISL's are considered to have variable length and each satellite decides on the neighboring satellite to find the shortest delay path. In this approach, a satellite may change its decision in case of excessive queue length; however, it is desired to avoid congestion before it happens.

Since satellites cover smaller areas in low orbit systems, the traffic requirements are unbalanced due to high population in cities, low at rural areas and almost no population over the oceans which form 75 per cent of the Earth's surface. This may lead to congestion in some resources, while others are under-utilized. To overcome this problem, the routing algorithm should distribute the flows in balanced way over appropriate ISLs between the communicating nodes. Considering this issue, Explicit Load Balancing (ELB) is proposed in [42]. ELB scheme is based on traffic load information at the next hop satellite on the remainder of the path to destination. A congested satellite sends signals to its neighboring satellites to decrease their sending rates and its neighbors search for alternative paths in order to pass the extra burden to less congested satellites. However, ELB does not give any solution for the case where alternative paths are also congested. Moreover it does not take any action for load balancing until some nodes experience a level of congestion. There are also some flow-based routing algorithms that aim balanced distribution of traffic flow to the network resources. Maximum-Flow Minimum-Residual (MFMR) algorithm proposed in [53] is a good example for the flow-based routing algorithms. MFMR tries to minimize maximum flow over a given set of shortest paths and hence avoid congestion by achieving balanced distribution of traffic. The main drawback of MFMR algorithm is that it implies knowledge of the flows over these paths and it does not consider fast dynamic changes in

the traffic flow over the given paths. In [40], Papapetrou *et al* propose an Adaptive Flow Deviation algorithm which aims to balance the traffic load via a flow deviation algorithm. This algorithm has also similar drawbacks as MFMR, and it implies high signaling overhead and high complexity with no guarantee of performing better than the simpler shortest path routing algorithms.

In this chapter, we propose an adaptive routing algorithm for satellite networks, namely Priority-based Adaptive Routing (PAR) algorithm, which distributedly sets the minimum-hop path towards a destination, and is more suitable for dynamic traffic. PAR takes the past utilization and queuing information of links into account and aims to achieve more uniform load distribution. In addition, we make an enhancement on PAR for better utilization of the ISLs and propose ePAR. Using simulation results we show that the proposed techniques not only increase throughput but also decrease delay. Furthermore, in order to further increase the performance of the system, we propose and evaluate a deflection routing mechanism, which deflects the packets to longer routes when the outgoing links in shortest paths are not available. Finally, we present a detailed analysis of ePAR. Since there is a number of parameters that should be adjusted properly, our analysis provides an opinion on the setting of these parameters.

4.2. Proposed Adaptive Routing Algorithms

In the context of satellite constellations, we can define “shortest path” in two ways: “minimum hop path” and “minimum delay path”. In the former one, we do not consider the dynamic length changes in ISLs and assume that ISLs are of fixed length. The shortest path is the one that passes minimum number of hops. In the latter, we consider that the distance measure is the total propagation delay and the length changes in ISLs should be taken into account. In the literature, some of the proposed algorithms consider that a route that traverses less satellite nodes is shorter and try to minimize average hop-count per transmission, while some others aim to minimize average end-to-end delay. In this work, we consider the former case, i.e. our proposed adaptive routing algorithms aim to use shorter paths in terms of hop-count. This is a reasonable assumption for most of the

constellation topologies. However, we also note that our proposed algorithms can easily be adapted to "minimum-delay path" case. We will examine this case in Section 4.2.4.

4.2.1. Priority-based Adaptive Routing

Satellite networks are usually modeled as regular mesh-like networks. In a regular mesh-like network, there might be many shortest paths between a source-destination (s-d) pair, in terms of hop-count. At each satellite node, more than one outgoing link could be on one of the minimum hop path. Decision on sending the data from which of those links has an important effect on the distribution of the traffic and utilization of ISLs. In the proposed algorithm, the link to be used is decided by a priority mechanism depending on the past utilization information about the links. We call this technique Priority-based Adaptive Routing (PAR). The priority metric used for this purpose can be determined in various ways. One simple possibility is to set it to the number of packets arrive to the link. However, the congestion in a link is not only related with the number of arrivals to it. For example among the links with same number of arrivals, the link with shorter queue length may be favorable. Therefore we support the idea that the priority metric should include past information for both utilization and length of the queues. While other functions can be investigated to select the best priority metric, for sake of simplicity and clarity we use linear combination of utilization and average queue length as follows:

$$\mu = \alpha \cdot n_t + \beta \cdot l_q \tag{4.1}$$

where n_t is the successfully transmitted data per second in the corresponding link, and l_q is the average queue length. Each link has its own μ value, and it is updated depending on the changes in the traffic. Using this metric, traffic tends to distribute the links in a more balanced way. Note that α and β are design parameters that should be adjusted properly due to the traffic requirements and network topology.

4.2.2. Enhanced Priority-based Adaptive Routing

It is important to note that most of the contentions occur between packets with different source-destination (*s-d*) pairs. Moreover, it is better to transmit packets of a particular flow over a single route, in order to avoid packet reordering and delay jitter. Therefore it would be better to switch packets with same *s-d* pairs to the same outgoing link. This suggests that the performance of PAR algorithm may be enhanced by using the following metric:

$$\mu = \alpha \cdot (n_{\mathrm{t}} - n_{\mathrm{t}}^{sd}) + \beta \cdot l_{\mathrm{q}} \tag{4.2}$$

where n_{t}^{sd} is the amount of transmitted packets corresponding to the *s-d* route, and μ_{sd} is the priority metric for traffic traversing on the *s-d* route. At the expense of increased complexity on satellite nodes, better ISL utilization may be achieved by this technique. This technique is called enhanced PAR (ePAR). We will describe results obtained by simulation for a considered constellation topology and how the adjustment of α and β parameters in ePAR could affect the performance of the system, in the following sections.

4.2.3. Aging Mechanism

Considering that the latest utilization and buffering information is more important than the older ones, an aging mechanism is needed while computing the priority metric. One possibility is to take the average of the last *t* seconds. However this mechanism has some drawbacks. Firstly, the information belonging to earlier times are also important and ignoring them completely is not reasonable. Moreover, storing the information for the last *t* seconds involves increased memory complexity. Therefore, we propose an aging mechanism as follows.

We define an aging period with length t_{a}. At the beginning of each period, we store the current μ value in a variable called μ^{o}. Then satellite starts to collect utilization and buffering information in a new variable called μ^{n}. At t_0'th time unit of a given period, μ is calculated as follows:

$$\mu = \mu^{o} \cdot \left(1 - \frac{t_0}{2t_a}\right) + \mu^{n} \cdot \left(\frac{t_0}{2t_a}\right) \tag{4.3}$$

Equation 4.3 is for PAR. For ePAR, it can be rewritten as:

$$\mu_{sd} = \mu_{sd}^{o} \cdot \left(1 - \frac{t_0}{2t_a}\right) + \mu_{sd}^{o} \cdot \left(\frac{t_0}{2t_a}\right) \tag{4.4}$$

PAR and ePAR does not have any signaling complexity because each node uses only local information. They have small amount of computational complexity and space complexity for calculating, storing and aging the priority metrics. However, ePAR needs extra complexity for taking care about information of particular flows. If there are too many source-destination pairs in the network, then ePAR may become infeasible.

4.2.4. Priority-based Adaptive Minimum-Delay Path Routing

In this work, we consider a priority-based adaptive routing for satellite networks, aiming to minimize hop-count. However, for some constellations, it could be more appropriate to consider the amount of propagation delay instead of the hop count. In fact, one can also argue that better definition of the "shortest path" concept involves the sum of the total propagation delay and the expected queuing delay (which is also related to the number of hops) that a packet would experience from source to destination [54]. Since actual lengths of ISLs are not identical and they dynamically change depending on the movement of satellites, it is somewhat challenging to design a routing algorithm that cares about propagation delay. In this sub-section, we will examine how PAR can be extended for this purpose, and define PAR for Minimum Delay Path Routing (PAR-MD).

As it was mentioned before, dynamic changes in the ISL lengths are predictable and periodic due to the strict orbital movements of satellites. Therefore, some techniques are proposed in order to use this periodicity property of the dynamic topology. Virtual Node (VN) [9] and DVTR protocols are the most common ones. In VN technique, a fixed virtual topology consisting of virtual nodes is superimposed over the physical topology in order to

hide the mobility of satellites from routing protocols. Each satellite corresponds to a VN at any given time. As a satellite disappears over the horizon, its corresponding VN becomes represented by the next satellite passing overhead and the state information (such as routing table entries) is transferred to it. VN is not appropriate to use with PAR, because PAR uses the utilization and buffering information of the physical satellites, not their corresponding virtual nodes. Therefore, we will not deal with this technique in the rest of this chapter. On the other hand DVTR technique, as mentioned before, divides the system period into N time intervals. During an interval i, the topology is modeled as constant graph G_i. Time intervals are short enough to define length of ISLs as fixed. Then the shortest paths and alternative paths for each time interval can be set by using well-known algorithms like Dijkstra's shortest path algorithm. Then PAR-MD can use these routes in order to find the most appropriate outgoing link of a given intermediate satellite, for a given *s-d* traffic. We said that there could be many minimum hop paths between a *s-d* pair in a satellite network. Therefore, in an intermediate satellite node, PAR was selecting the outgoing link which was over a minimum hop path, and had the best priority. However this is not the case for minimum delay paths because of the spherical shape of the Earth. Therefore, we may consider k-shortest paths, or the paths that are not longer than the shortest path with a given degree, while selecting appropriate outgoing link. In other words, an intermediate node selects the outgoing link which is over one of these paths and has the best priority. In this case, priority metric of a link may also include the length of short paths that pass over it. However, this would require an extra complexity and overhead in satellite nodes.

4.2.5. Deflection Enabled PAR (DEPAR)

PAR uses only the ISLs that are on a minimum hop path. In the case that these paths are congested, it drops the packet and does not utilize other links. However, instead of dropping packets, it could be more appropriate to utilize a deflection routing mechanism, in other words to deflect packets to longer routes. In this sub-section, we define and describe a deflection routing strategy to use together with the PAR algorithm.

In PAR, each satellite forwards a packet to one of its neighbors that is on a minimum-hop path for the corresponding packet. Now, we define a deflection routing mechanism which will be used when all of the outgoing links towards those neighbors are congested. The proposed deflection routing algorithm is as follows:

When a satellite receives a packet (from a terrestrial node or a satellite node):

- It checks the outgoing ISLs that are included in one of the shortest paths from the source node of the packet to its destination. Let's say these ISLs, primary ISL. Among these ISLs, firstly it tries to send the packet from the link with highest priority. If it is congested, it tries other primary ISL(s), if there exists any.

- If all of the links over a minimum-hop path are congested, we select an ISL that is not on a shortest path. The link for deflection must be a neighboring link of one of the primary ISLs, and we call these ISLs, secondary ISL. For example, consider a constellation with 4 ISLs per satellite: West (*W*), East (*E*), North (*N*), South (*S*). For a particular packet, if *N* is the only primary ISL, *W* and *E* are the secondary ISLs. If *N* and *E* are the primary ISLs, *W* and *S* are the secondary ISLs. Among the secondary ISLs, decision of which ISL to deflect the packet depends on the same priority mechanism. If the secondary ISL with high priority is congested at that instant, then the one with low priority is selected. If that link is also congested, then packet is dropped.

- In the case of deflection, ID of the corresponding satellite is written over the packet, in order to prevent the packet to revisit that satellite. Otherwise, the routing algorithm will not be loop-free.

Another issue in the context of deflection routing is the threshold for number of deflections. If no threshold is defined, packets may waste resources unnecessarily. Therefore we propose to supply a threshold as follows: When a packet needs to be deflected, we account for the number of hops it has traversed so far. If it exceeds minimum hop distance between the source satellite node and the corresponding node, with a predetermined threshold, packet is dropped. Otherwise it is deflected. To formulate this,

we define h_i (number of hops packet p_i traversed so far), $mhd_{x,y}$ (minimum hop distance between satellite x and satellite y) and d (predetermined threshold). If s is the source node of the packet and c is the corresponding node, a packet could be deflected if the following situation holds

$$\{h_i < mhd_{s,c} \cdot d \quad | \quad h_i = 0\} \tag{4.5}$$

It is clear that none of the satellites (except the source node) support deflection routing at $d = 1$, and packets will always be deflected for large d values, if possible. Here, the question we tackle is "which d value should be set to improve system performance?" We investigate the answer in Section 4.4.4 for various traffic load characteristics after testing contribution of DEPAR via extensive set of simulations.

4.3. Satellite Network Architecture and Routing Details

4.3.1. Network Topology

In this work, we consider a polar LEO constellation similar to Teledesic, with 12 orbit planes and 24 satellites per plane at a height of 700 km. It is a π-constellation, where there is a seam between satellites moving in opposite direction. Figure 4.1 shows the considered network topology.

We assume that there is no ISL passing the seam. As shown in the Figure 4.1, seam divides the network into two parts and the satellites over the eastern hemisphere and the satellites over the western hemisphere move in opposite directions. Hence, a data that originates from a location at one hemisphere could be sent to a location in the other hemisphere, only by passing a pole. Although this is an important drawback of π-constellations; it is not a critical factor in dramatically affecting the performance of the proposed and tested routing techniques. In our topology, seam passes over the Pacific and Atlantic oceans as shown with bold lines in Figure 4.2. Due to complexity of the system parameters and to simplify the analysis we also assume that satellites have disjoint footprints and dividing the Earth into 12 × 24 terrestrial zones, as in Figure 4.2, and each

satellite sees one of these zones. Another assumption is made on in the handover mechanism, i.e., as the satellites move with angular velocity of 3.6 degree per minute, they switch their zones in a discrete manner. Each zone is represented by $Z_{x,y}$, where $x \in (0,11)$ and $y \in (0,23)$. x is the orbit plane number of satellites passing over that zone, and y is defined as follows. For western hemisphere, zones that are nearest to the northern pole have a y value of zero. Going to the south, y is incremented by one. At the eastern hemisphere y is 12 for the zones nearest to the southern pole and going to the north, it is incremented by one.

Although more realistic scenarios could have been selected in the simulations, the potential of our algorithms should remain the same.

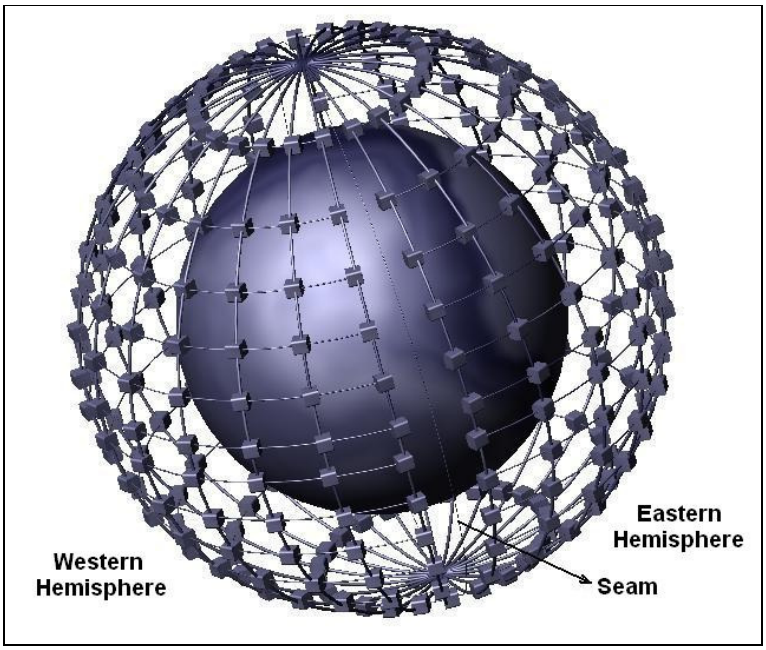
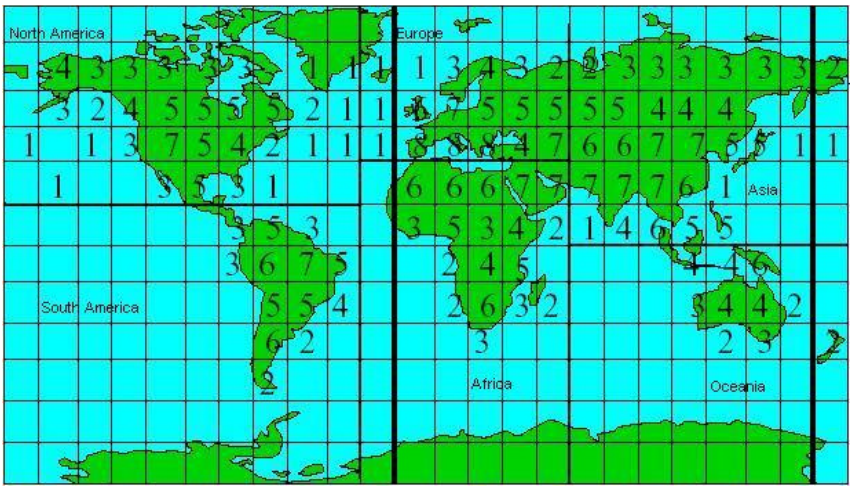


Figure 4.1. Polar π-constellation topology with 12×24 nodes



Intensity level	1	2	3	4	5	6	7	8
Traffic (million minutes per year)	1.6	6.4	16	32	95	191	239	318

Figure 4.2. Earth zone division, and user intensity levels on each zone (for year 2005 [55]).

4.3.2. Routing Details

In a mesh network topology as shown in Figure 4.1, there is more than one shortest path between each source-destination pair (except if they are in same latitude or longitude) in terms of hop-count. In the case of static routing, only one of these routes is utilized. If the adaptive route is only set in the source node, as in [52], this also does not yield a good utilization of ISLs. However, routing techniques which also employ intermediate nodes for route computation give better performance results. When a satellite receives a packet, it looks for its destination node. If it is in the same latitude or longitude, there is only one direction to send (for the shortest path). Otherwise, there are two possibilities. In that case, determining which direction to send depends on the routing algorithm. For this purpose, we define four different adaptive shortest path routing algorithms: Fixed Adaptive Routing (FAR), Random Adaptive Routing (RAR), Priority-based Adaptive Routing (PAR), and Enhanced Priority-based Adaptive Routing (ePAR). In this section, we first explain how to find the outgoing direction, and then clarify these routing techniques.

4.3.2.1. Direction Estimation. We define two variables: $dir_x \in$ {East, West}, and $dir_y \in$ {South, North}. Let's consider that a satellite node c, receives a packet with destination d. Assuming that c is over the zone $Z_{xc,yc}$, and d is over the zone $Z_{xd,yd}$, determination of dir_x

and dir_y on node *c* is done according to the pseudocode given in Figure 4.3. Note that, given direction estimation process is specific for the considered mesh-like topology given in Figure 4.1. For any other constellation topology, a specific direction estimation process should be determined depending on the minimum-hop paths between satellite nodes.

Following the determination of directions, the next task is to determine which route to select first. In this context, we may have various routing techniques.

```
if xc = xd
       dir_x = {}
else if xc < xd
       dir_x = East
else if xc > xd
       dir_x = West

if yc = yd
       dir_y = {};
else if yc < 12 AND yd < 12 {
       if yc < yd
              dir_y = South
       else
              dir_y = North
}
else if yc ≥ 12 AND yd >= 12 {
       if yc > yd
              dir_y = South
       else
              dir_y = North
}
else if yc < 12 AND yd ≥ 12 {
       if (yd-yc) < (24-yd+yc)
              dir_y = South
       else
              dir_y = North
}
else if yc ≥ 12 AND yd < 12 {
       if (yc-yd) < (24-yc+yd)
              dir_y = South
       else
              dir_y = North
}
```

Figure 4.3. Algorithm for determining directions towards a destination

4.3.2.2. Routing Algorithms. In what follows, we define one static and four new adaptive shortest path algorithms considered in this study:

1. Static Shortest Path Routing (STA): When a satellite receives a packet, it sends it in *y* direction (South or North) if the satellite is not in the same latitude with the destination. Otherwise, it sends in *x* direction (East or West). The established route consists of two

straight paths that first goes in *y* direction, and then in *x* direction (of course if *s* and *d* are in different longitudes and latitudes).

2. Fixed Adaptive Routing (FAR): A satellite that receives a packet, always selects dir_y as the initial direction. If dir_y is empty or ISL on that direction is busy, it tries dir_x.

3. Random Adaptive Routing (RAR): Satellite randomly selects one of the dir_y or dir_x. If it is empty or ISL on that direction is busy, it tries the other direction.

4. Priority-based Adaptive Routing (PAR): Satellite checks μ values for ISLs on both directions. It selects one with less μ value as initial direction. If the lSL on that direction is busy, it tries the other.

5. Enhanced PAR (ePAR): Satellite checks the μ_{sd} values for ISLs on both directions, where *s* is the source node and *d* is the destination node of the packet. It selects the direction of the ISL with less μ_{sd} value as initial direction. If that ISL is busy, it tries the other.

6. Deflection Enabled PAR (DEPAR): PAR with deflection routing that is described in Section 4.2.5.

7. Deflection Enabled FAR (DEFAR): FAR with deflection routing. Same deflection algorithm is used as DEPAR. However, among the secondary ISLs, which link to deflect first is decided randomly. If the decided ISL is congested at that instant, then the other secondary ISL is selected. If that link is also congested, then packet is dropped.

In all cases, we assume that *"the ISL is busy"* means it is transmitting a packet and its buffer is full at that moment. Depending on the network characteristics, one may prefer to set a threshold value for the buffer size, and consider the ISL as busy if its queue length exceeds this threshold value. This is desirable especially for ISLs with high buffer capacities.

4.3.2.3. Contention Resolution Techniques. Some contention resolution schemes are already defined in the literature for the situations that two packets arrive to an ISL at the same time [37]. Random Packet Win (RPW), Oldest Packet Win (OPW), and Shortest Hop Win (SHW) are the most common contention resolution techniques. In this work, we assume that SHW is utilized (except the simulations described in Section 4.4.4, where RPW is utilized). SHW favors the packets with the shortest hop distance to its destination node, and RPW chooses one of the contending packets randomly.

4.4. Simulations

4.4.1. Simulation Setup

To test the performances of the proposed algorithms, we use an extensive set of simulations. Simulation scenarios and system parameters are chosen to highlight the algorithm's capability. We simulate the constellation shown in Figure 4.1. It is a polar π-constellation with 12 × 24 satellites, where there exists a seam. Satellites that are in the border of seam have three ISLs, since we assume that there is no ISL over seam. Every other satellite has four ISLs. All satellites rotate on their orbit with an angular velocity of 3.6 degrees per minute. This means that their corresponding terrestrial zone changes at each 250 seconds. They complete their rotation in 100 minutes. For simplicity, all ISLs are assumed to be identical (in terms of length and capacity) and their capacity is assumed to be 0.16 Gbps. Each ISL has a buffer of size 40 Mbytes. A packet size is assumed to be 1 Kbyte. Therefore, ISL capacity and buffer size are considered as 20,000, and 40,000 packets, respectively.

4.4.2. Traffic Model

Our traffic model is similar to the model considered in [45]. It depends on the 2005 statistics about the user density levels per zone (Figure 4.2), Internet host density levels per continent (Table 4.1), and user activity levels per hour in percentage of the total traffic (Figure 4.4).

Table 4.1. Internet Host Distribution by Continent (January 2005) [56]

Continent	Number of Hosts (h_C) (x10^3)	Percentage
North America	223545,1	70,45
Europe	52947,1	16,69
Asia	28511,4	8,98
South America	6026,2	1,9
Oceania	5621,6	1,77
Africa	671,3	0,21

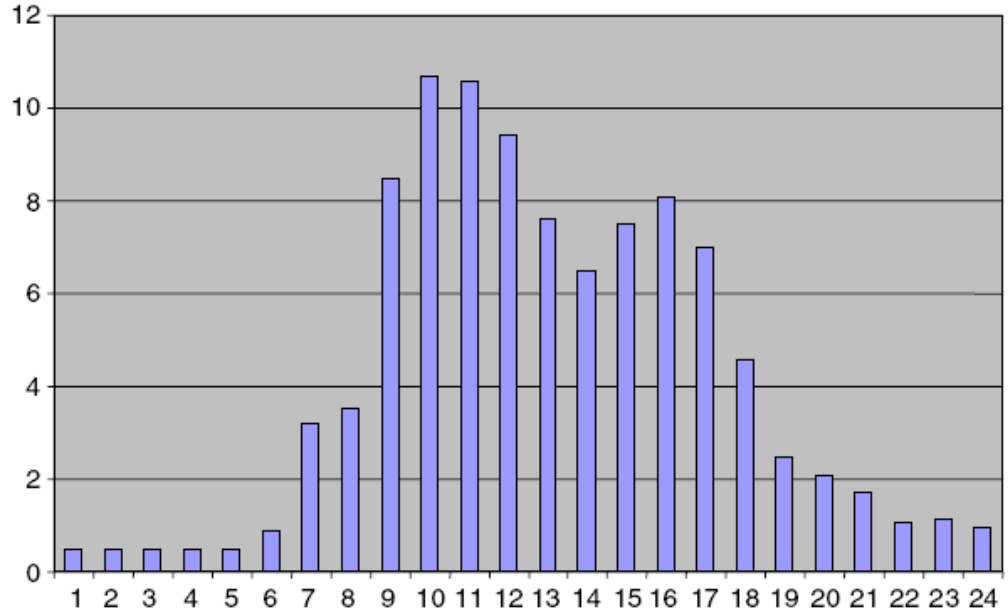


Figure 4.4. User activity percentage per hour [57]

Let $u_{x,y}$ be the user density of zone $Z_{x,y}$. We set the host density level of zone $Z_{x,y}$, $h_{x,y}$, as the portion of total host density of its continent (C_i) that is proportional with its user density:

$$h_{x,y} = h_{C_i} \cdot \frac{u_{x,y}}{\sum_{Z_{a,b} \in C_i} u_{a,b}} \tag{4.6}$$

[58] suggests a traffic generating method depending on these densities. Traffic requirement from zone $Z_{x,y}$ to zone $Z_{t,k}$, $T(Z_{x,y}, Z_{t,k})$, is proportional with the user density in $Z_{x,y}$, $u_{x,y}$, host density in $Z_{t,k}$, $h_{t,k}$, and distance between these two zones ($dist(Z_{x,y}, Z_{t,k})$):

$$T(Z_{x,y}, Z_{t,k}) = \frac{\left(u_{x,y} h_{t,k}\right)^{\theta}}{\left(dist\left(Z_{x,y}, Z_{t,k}\right)\right)^{\psi}} \tag{4.7}$$

In the simulations, we set $\theta = 0.5$ and $\psi = 1.5$ (as in [45]). Depending on this traffic requirement matrix, we model the traffic. We assume that, at a given hour *h*, the arrival of a packet with source = $Z_{x,y}$ and destination = $Z_{t,k}$ is a poisson process with rate $\lambda(Z_{x,y}, Z_{t,k}, h)$ packets/second:

$$\lambda(Z_{x,y}, Z_{t,k}, h) = \frac{T(Z_{x,y}, Z_{t,k})}{\sum_{\forall Z_{a,b}} \sum_{\forall Z_{c,d}} T(Z_{a,b}, Z_{c,d})} \cdot \frac{a_h}{100} \cdot \frac{A}{3600} \tag{4.8}$$

where, *h* is the current local hour and a_h is the activity percentage in the corresponding hour (*h*), that is given in Figure 4.4. Moreover, *A* is the aggregate traffic that represents total traffic generated worldwide (packets per day).

4.4.3. Simulation Results

We implemented all routing techniques based on the described network topology and the traffic model given above. We developed our own simulator in C++. We tested the performance of routing algorithms in terms of drop ratio and average queue length per link. Drop ratio is defined as the ratio of dropped packets to the sum of dropped and successfully transmitted packets, and average queue length is the ratio of the sum of the average number of packets in all buffers to the number of ISLs.

We set the system parameters to the values shown in Table 4.2. Best selection of the α and β values depends on the traffic, and network characteristics. In Section 4.5, we include some analysis in order to find a way on how to adjust these parameters. In the simulations, in order to avoid complication, we simply decide to set these parameters as given in Table 4.2, so that $\alpha \cdot n_t$ ranges between zero and one, and $\beta \cdot l_q$ ranges between 0 and 2. Note that, in parameter selection for ePAR algorithm, analysis provided in Section 4.5 could be utilized for improving system performance.

Table 4.2. System parameters

Total simulation time	1 day
Warm-up period	60 seconds
Aging period (t_a)	25 seconds
α	0.00005
β	0.00005

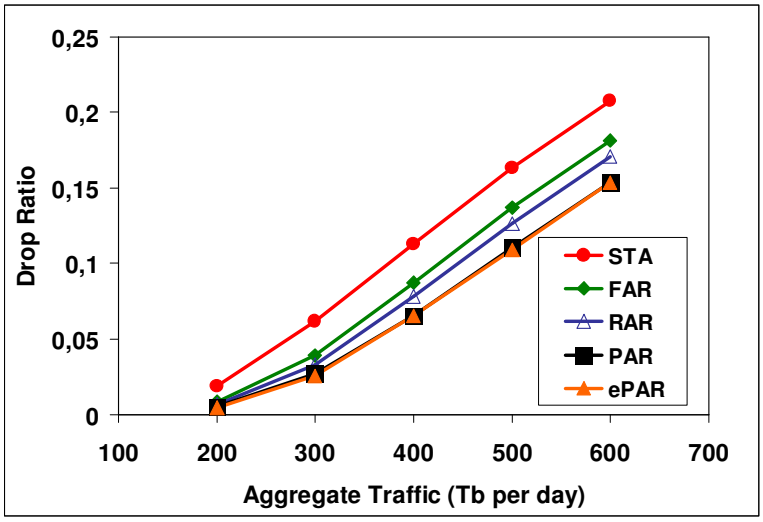


Figure 4.5. Drop ratio versus aggregate traffic for five different routing techniques

Figure 4.5 shows the drop ratio versus A (in terms of terabit per day). As expected, Static Routing performs the worst. FAR never provides a balanced distribution of traffic, therefore its performance is worse than other adaptive routing techniques. It can be seen that priority-based algorithms are the best in case of drop ratio. An important observation is that there is no valuable difference between the performances of PAR and ePAR. This could be because the fact that there are too many nodes, and hence too many source-destination pairs. In this case the significance of channeling packets with same *s-d* pairs to same links is not evident. Moreover, as number of nodes increase, the complexity of ePAR increases. Therefore, for the networks with large number of nodes, PAR seems to be more suitable technique than ePAR. This suggests that ePAR should be further investigated for MEO satellites.

Figure 4.6 illustrates the difference between queue lengths for different routing schemes. Static Routing has the least queue lengths since most of the packets are dropped

without being buffered. Because, there is no alternative route for Static Routing; hence packets should not have to wait anymore, if the link on the static route is busy. As the number of successfully transmitted packets increase, we expect that the lengths of queues also increase because of the high utilization of links (obtained result that illustrates the difference between queue lengths for RAR and FAR meets our expectation). However, Figure 6 suggests that this is not the case for PAR and ePAR, and they outperform all other adaptive routing techniques in terms of queue length. This is because priority-based techniques provide balanced distribution of traffic among links, and more packets are successfully transmitted with less waiting times in queues. This means that proposed priority-based adaptive routing schemes decrease end-to-end delay, while increasing throughput.

Furthermore, we examine the performance behavior in hour base. Figure 4.7 illustrates generated traffic per hour (MAX), and successfully transmitted data (in Terabits) for each routing algorithm. Results are for A = 400 Tbps. The base time is Greenwich Mean Time (GMT). Two peaks are observed in the number of generated data. The first peak corresponds to the time when it is daytime and user activities are in peak levels in Europe, and other corresponds to the time when activities speed up in Northern America. In the second peak time, performance difference between routing algorithms are more evident, whereas in the first peak time, all adaptive routing algorithms perform similar. The reason for this may be due to the traffic model. That is, for packets originating from Europe, there exist some factors that cause packet drops regardless of which routing algorithm is used. For example, we observe that most of the packets drops occur in the first hop. In other words, most of the packets received from the terrestrial transmitters could not be passed to neighboring satellite, since links in both directions are busy. This condition could not be resolved by any shortest path routing algorithm. Deflection Enabled PAR algorithm could be utilized to overcome those cases.

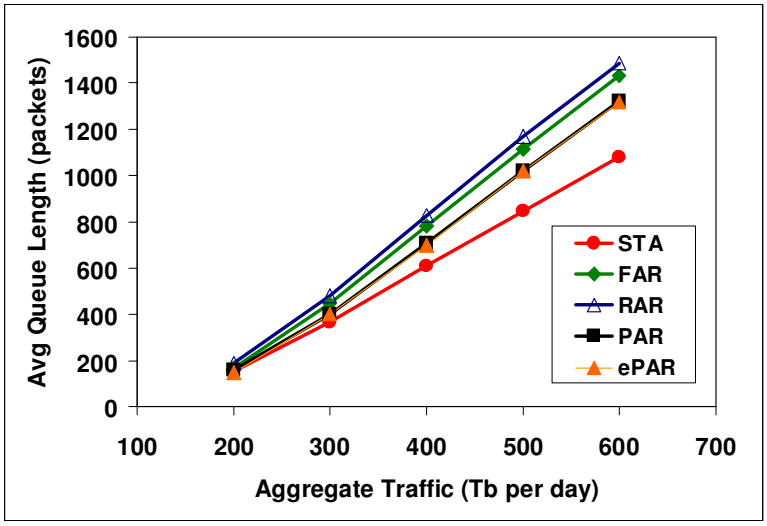


Figure 4.6. Average queue length versus aggregate traffic

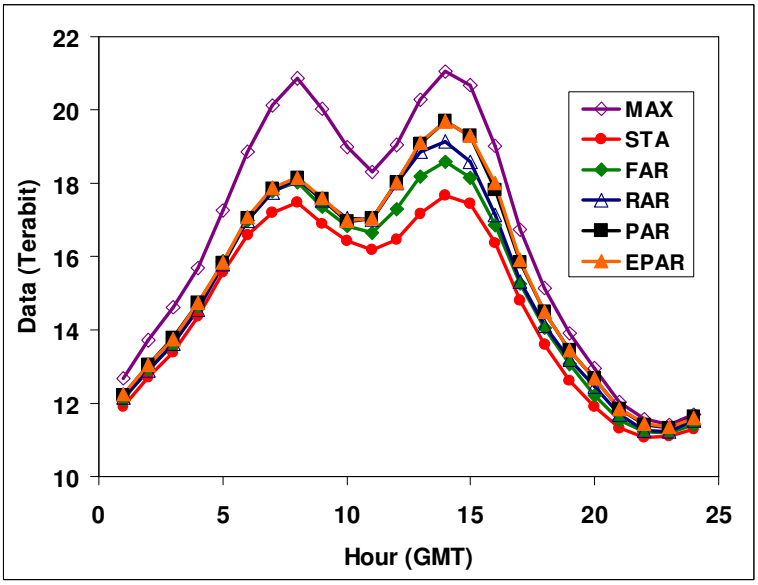


Figure 4.7. Successfully transmitted data (Tb) versus hour (GMT)

4.4.4. Contribution of DEPAR

In order to test the contribution of deflection routing, we evaluate the performance of DEPAR algorithm. Figure 4.8 shows the drop ratio versus *A* (in terms of terabit per day). For DEPAR and DEFAR we set the *d* value to 1.2. It is evident that proactive priority based algorithms (PAR and DEPAR) outperform non-proactive algorithms (FAR and DEFAR). More interesting observation is on the performance difference between

deflection routing enabled algorithms (DEPAR and DEFAR) and their pure versions (PAR and FAR). For low traffic loads, deflection routing enabled algorithms perform better. For example, for A = 200 Tbps, drop ratio for DEPAR is 1/18 of the drop ratio for the PAR algorithm (although it is not visible in the Figure 4.8). For A = 300 Tbps, this ratio becomes approximately 1/3. As the traffic load increases, difference between drop ratios of DEPAR and PAR reduces. For A = 800 Tbps, performances of two algorithms seem to be same in terms of drop ratio. Same relation is also valid for FAR and DEFAR. While for low traffic loads DEFAR outperforms FAR, for high traffic loads the situation is reversed. This is because deflection enabled algorithms postpone dropping of packets and further increase the traffic load in the system. Packets traverse more hops in the system, but because of the high traffic load, number of packets that reach to destination reduces. For low A values, system can tolerate the extra load caused by the deflection mechanism.

Figure 4.9 illustrates the differences between queue lengths for different routing schemes. As the number of successfully transmitted packets increase, we expect that lengths of queues also increase because of the high utilization of links. However, this is not the case for PAR, and it outperforms fixed adaptive routing scheme (FAR) in terms of average queue length. This is because priority-based techniques provide balanced distribution of traffic among links, and more packets are successfully transmitted with less waiting times in queues. Same comparison is also true between DEPAR and DEFAR. However, according to the obtained results, queue length values for deflection enabled algorithms are worse than that for pure versions. This is because in deflection enabled algorithms, packets stay in the system for longer times. Therefore traffic load is increased, and this results in more waiting times in buffers. However, we ignored retransmissions in our simulations. Algorithms without deflection cause higher packet drops for low aggregate traffic load. This means that higher number of retransmissions is needed. Therefore we can say that, if the retransmissions were taken into account, average queue length values for deflection enabled algorithms and their pure versions would be closer to each other.

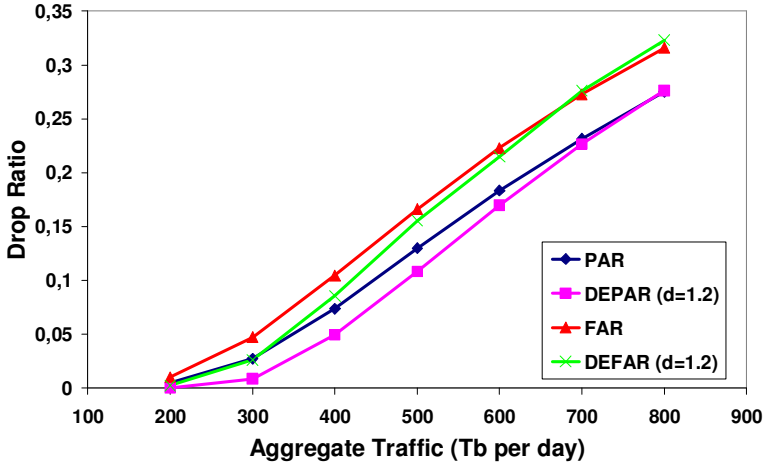


Figure 4.8. Drop Ratio versus Aggregate Traffic

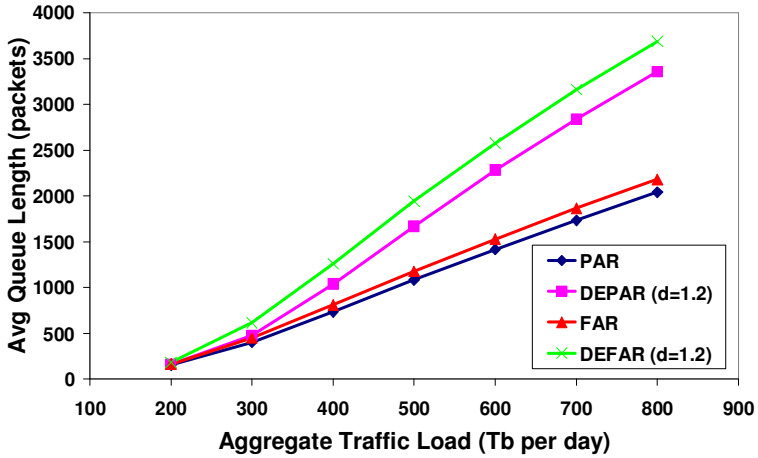


Figure 4.9. Average Queue Length versus Aggregate Traffic

Next, we examine how deflection routing affects the length of path per packet. According to Figure 4.10, packets traverse more hops to reach its destination in deflection routing enabled algorithms. This is an expected result because PAR and FAR are shortest path algorithms, whereas deflection routing also utilize longer paths. For PAR and FAR, average hop count decreases as traffic load increases, because packets belonging to long distant routes are exposed to more drops for crowded systems. For deflection enabled algorithms, average hop count per successfully transmitted packets increase with the traffic load up to some point, and then it start to decrease. This could be explained as follows: For very low traffic levels, fewer packets are exposed to deflection. Therefore average hop count is less. As traffic load increases, more packets will be deflected and average hop

count increases. However, after a point, crowdedness of the system leads to higher drop ratio for long distance dependent traffic and therefore number of hops traversed per packet starts decreasing.

Obtained results show that for low-to-moderate traffic loads, DEPAR yields better throughput with a reasonable increment in the delay. However, as traffic load increases, pure PAR algorithm becomes more advantageous than DEPAR.

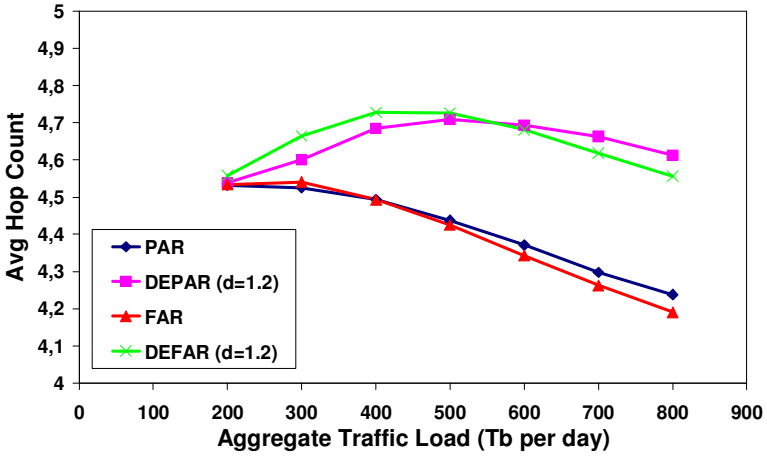


Figure 4.10. Average Hop Count versus Aggregate Traffic

Until this point we set *d* value for DEPAR to 1.2. However, the effect of the *d* value should be investigated for improved system performance (throughput, delay, etc) under various traffic conditions. For this purpose we run our simulations for various *d* values under various traffic loads.

Figure 4.11 suggests that drop ratio is not much affected from the *d* value. We think this is because links near to the high-traffic-generating source nodes are generally more crowded and therefore most of the drops occur in initial hops. Therefore increasing *d* value has very small contribution on the throughput of the system. In DEPAR with d=1.0, deflection is allowed only in the source node and therefore perform worse than cases with slightly more *d* values.

Although the throughput is not much affected from the *d* value, this is not the case for queue length and average hop count traversed per packet. As shown in Figure 4.12 and Figure 4.13, there is a perceptible effect of *d* value on queue length and number of hop traversed. These observations suggest that when we use DEPAR, it is reasonable to keep *d* value in low levels. However, we should note that these results are obtained for our simulation model (with the given network topology, given traffic model, etc.). Therefore, results may change for different topology and traffic generation characteristics.

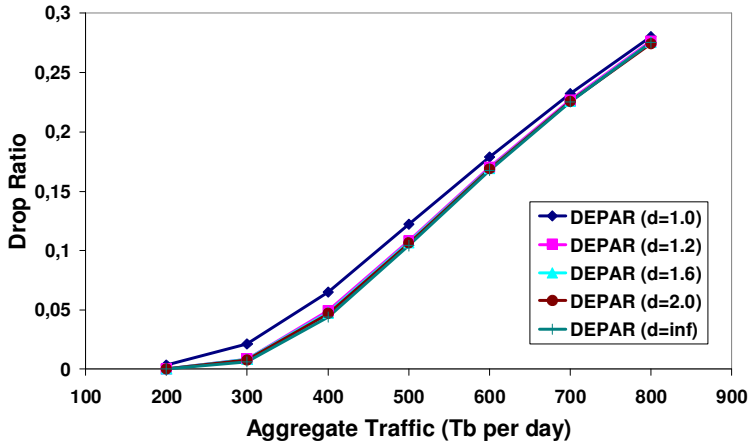


Figure 4.11. Drop Ratio versus Aggregate Traffic for various *d* values

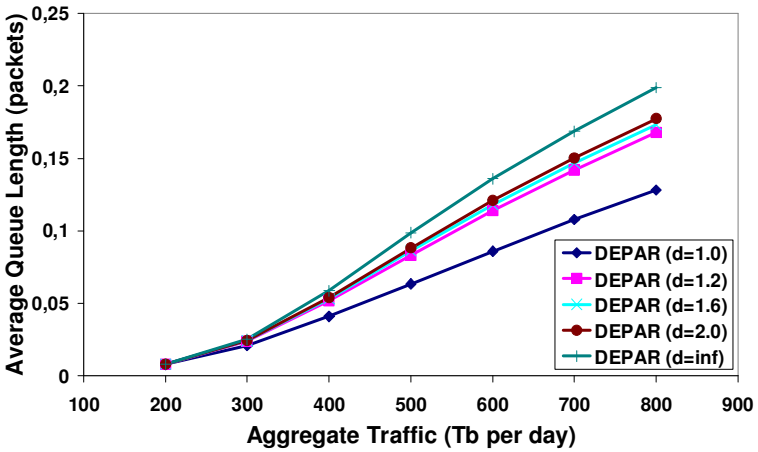


Figure 4.12. Average Queue Length versus Aggregate Traffic for various *d* values

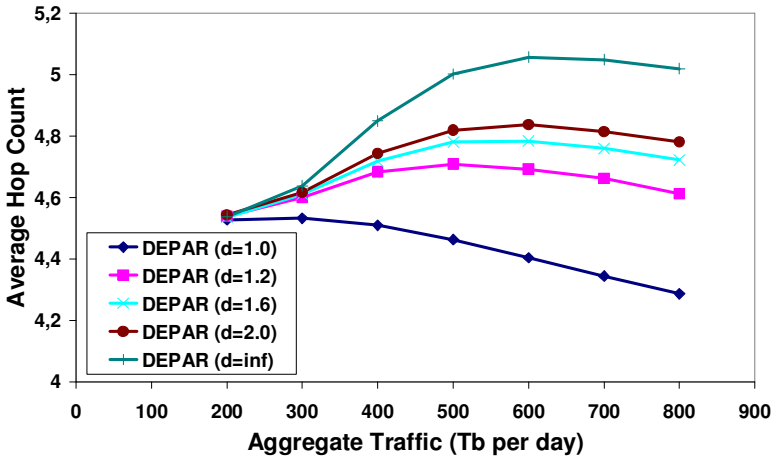


Figure 4.13. Average Hop Count versus Aggregate Traffic for various *d* values.

4.5. Parameter Selection for Increased Stability

This section includes an analysis of the ePAR algorithm, utilizing the mentioned aging mechanism. Consider that α and β in Equation 4.2 denote the design parameters that should be adjusted properly depending on the network characteristics. By the optimal selection of these parameters, not only better load distribution can be achieved, but also traffic flows can be made more stable. By stability, we mean avoiding the needless fluctuations due to redirection of all the flows in a congested link, simultaneously. In this section, we will investigate how to set these parameters to achieve more stable systems.

Defining a new variable $n_{sd} = n_{\mathrm{t}} - n_{\mathrm{t}}^{sd}$, Equation 4.2 can be reduced to:

$$\mu_{sd} = \alpha \cdot n_{sd} - \beta \cdot l_{\mathrm{q}} \tag{4.9}$$

Note that, considering the aging mechanism in Equation 4.4, n_{sd} and l_{q} can be calculated in same manner as μ_{sd}:

$$n_{sd} = n_{sd}^{\mathrm{o}} \cdot \left(1 - \frac{t_0}{2t_{\mathrm{a}}}\right) + n_{sd}^{\mathrm{n}} \cdot \left(\frac{t_0}{2t_{\mathrm{a}}}\right) \tag{4.10}$$

$$l_{\mathrm{q}} = l_{\mathrm{q}}^{\mathrm{o}} \cdot \left(1 - \frac{t_0}{2t_{\mathrm{a}}}\right) + l_{\mathrm{q}}^{\mathrm{n}} \cdot \left(\frac{t_0}{2t_{\mathrm{a}}}\right) \tag{4.11}$$

Suppose a satellite network in which arrival rate of an *s-d* flow is Poisson distributed with mean $1/\lambda$ bps. We represent a flow with source *x*, and destination *y* with f_{xy}. We consider a scenario, where a new flow is participated to a link that is already utilized just below its capacity, and the data arrival rate exceeds its capacity by the participation of the new flow.

We assume that the existing aggregate flow (F_{e}) in a particular link has a rate of X/λ bps. A new flow (f_{mn}) with rate r_{mn}/λ is participated to the corresponding link. Note that the source destination pair of the new flow is $\{m,n\}$. After the aggregation of the existing flow F_{e}, and the new flow f_{mn}, the total arriving flow becomes $(X + r_{mn})/\lambda$. Supposing that this value is greater than the link bandwidth, B/λ, total amount of traffic to serve is B/λ and total flow to be blocked per second is $(X + r_{mn} - B)/\lambda$. Assuming that the blocking probability for each flow is same, blocked portion of F_{e} will be:

$$b_F = (X + r_{mn} - B) \cdot \frac{1}{\lambda} \cdot \left(\frac{X}{X + r_{mn}}\right) \tag{4.12}$$

And the transmitted portion of F_{e} will be:

$$t_F = \frac{B}{\lambda} \cdot \left(\frac{X}{X + r_{mn}}\right) \tag{4.13}$$

For an *s-d* flow (a new flow or an existing flow), let us represent initial value of the n_{sd} (before the participation of new flow) with ψ_{sd}^{o}, and the value that n_{sd} will converge is represented with ψ_{sd}.

Since t_F is the total transmitted data per second except the portion corresponding to f_{mn}, it is equal to ψ_{mn}, the new value that n_{mn} will converge.

$$\psi_{mn} = \frac{B}{\lambda} \cdot \left(\frac{X}{X + r_{mn}} \right) \tag{4.14}$$

Before the participation to the corresponding link, n_{mn} was equal to the total transmitted data in the link, since n_{t}^{mn} was zero. We represent the old value of n_{mn} with ψ_{mn}^{o}:

$$\psi_{mn}^{o} = \frac{X}{\lambda} \tag{4.15}$$

For any other flow f_{ij} that uses the same link (and hence that is a part of F_{e}):

$$\psi_{ij} = \frac{B}{\lambda} \cdot \left(\frac{X + r_{mn} - r_{ij}}{X + r_{mn}} \right) \tag{4.16}$$

where ψ_{ij} is the new value that n_{ij} will converge. The initial value for n_{ij} is determined by the difference between total initial flow and the flow with source i and destination j:

$$\psi_{ij}^{o} = \frac{X - r_{ij}}{\lambda} \tag{4.17}$$

Amount of blocked data continuously increases as time passes, and t seconds after the participation of the new flow, length of queue becomes $t \cdot \frac{X + r_{mn} - B}{\lambda}$, assuming that the buffer is initially empty. Since we assume increment in queue length is linear, average queue length between time t_1 and t_2 is:

$$\frac{t_1 + t_2}{2} \cdot \frac{X + r_{mn} - B}{\lambda}$$

For clear illustration, we represent increment in the queue length (per second) with a new variable ξ:

$$\xi = \frac{X + r_{mn} - B}{\lambda} \tag{4.18}$$

Table 4.3. Changes in utilization and buffering information with the time

Initially	$n_{sd} = \psi_{sd}^{o}$ $l_{\mathrm{q}} = 0$
After t_1 secs	$n_{sd}^{o} = \psi_{sd}^{o}$; $n_{sd}^{n} = \psi_{sd}$, hence $n_{sd} = \psi_{sd}^{\mathrm{o}} \cdot \left(1 - \frac{t_1}{2t_{\mathrm{a}}}\right) + \psi_{sd} \cdot \left(\frac{t_1}{2t_{\mathrm{a}}}\right)$ $l_{\mathrm{q}}^{\mathrm{o}} = 0$; $l_{\mathrm{q}}^{\mathrm{n}} = \frac{t_1}{2} \cdot \xi$, hence $l_{\mathrm{q}}^{\mathrm{n}} = \frac{t_1}{2} \cdot \xi \cdot \left(\frac{t_1}{2t_{\mathrm{a}}}\right)$
After t_{a} secs	$n_{sd} = \psi_{sd}^{o} \cdot \frac{1}{2} + \psi_{sd} \cdot \frac{1}{2}$ $l_{\mathrm{q}} = \frac{t_{\mathrm{a}}}{2^2} \cdot \xi$
After $t_{\mathrm{a}}+t_1$ secs	$n_{sd}^{o} = \psi_{sd}^{o} \cdot \frac{1}{2} + \psi_{sd} \cdot \frac{1}{2}$; $n_{sd}^{n} = \psi_{sd}$, hence $n_{sd} = \psi_{sd}^{\mathrm{o}} \cdot \frac{1}{2} \cdot \left(1 - \frac{t_1}{2t_{\mathrm{a}}}\right) + \psi_{sd} \cdot \left(\frac{1}{2} + \frac{1}{2} \cdot \frac{t_1}{2t_{\mathrm{a}}}\right)$ $l_{\mathrm{q}}^{\mathrm{o}} = \frac{t_{\mathrm{a}}}{4} \cdot \xi$; $l_{\mathrm{q}}^{\mathrm{n}} = \frac{2t_{\mathrm{a}} + t_1}{2} \cdot \xi$, hence $l_{\mathrm{q}} = \left(\frac{t_{\mathrm{a}}}{4} \cdot \left(1 - \frac{t_1}{2t_{\mathrm{a}}}\right) + \left(\frac{2t_{\mathrm{a}} + t_1}{2}\right) \cdot \frac{t_1}{2t_{\mathrm{a}}}\right) \cdot \xi$
After $2t_{\mathrm{a}}$ secs	$n_{sd} = \psi_{sd}^{o} \cdot \frac{1}{2^2} + \psi_{sd} \cdot \left(1 - \frac{1}{2^2}\right)$ $l_{\mathrm{q}} = \left(t_{\mathrm{a}} \cdot \frac{1}{2^3} + 3t_{\mathrm{a}} \cdot \frac{1}{2^2}\right) \cdot \xi$
After mt_{a} secs	$n_{sd} = \psi_{sd}^{o} \cdot \frac{1}{2^m} + \psi_{sd} \cdot \left(1 - \frac{1}{2^m}\right)$ $l_{\mathrm{q}} = \left(t_{\mathrm{a}} \cdot \frac{1}{2^{m+1}} + 3t_{\mathrm{a}} \cdot \frac{1}{2^m} + 5t_{\mathrm{a}} \cdot \frac{1}{2^{m-1}} + \cdots + (2m-1) \cdot t_{\mathrm{a}} \cdot \frac{1}{2^2}\right) \cdot \xi = \left(\frac{t_{\mathrm{a}}}{4} \cdot \sum_{i=1}^{m} \frac{2i-1}{2^{m-i}}\right) \cdot \xi$
After $mt_{\mathrm{a}}+t_1$ secs	$n_{sd} = \psi_{sd}^{\mathrm{o}} \cdot \frac{1}{2^m} \cdot \left(1 - \frac{t_1}{2t_{\mathrm{a}}}\right) + \psi_{sd} \cdot \left(\frac{1}{2^m} + \frac{1}{2^m} \cdot \frac{t_1}{2t_{\mathrm{a}}}\right)$ (4.19) $l_{\mathrm{q}} = \left(\left(\frac{t_{\mathrm{a}}}{4} \cdot \sum_{\mathrm{i}=1}^{m} \frac{2i-1}{2^{m-i}}\right) \cdot \left(1 - \frac{t_1}{2t_{\mathrm{a}}}\right) + \left(\frac{2mt_{\mathrm{a}} + t_1}{2}\right) \cdot \frac{t_1}{2t_{\mathrm{a}}}\right) \cdot \xi$ (4.20)

Table 4.3 shows how n_{sd} value and the queue length information changes after the participation of new flow, f_{mn} to the existing flow F_{e} in a particular link. The effect of aging mechanism is also considered. Let $t_1 < t_{\mathrm{a}}$ where t_{a} denotes the length of aging period.

Without loss of generality, we consider that the participation of new flow occurred at the beginning of an aging period. Note that, the equations given in Table 4.3 are valid for both the new flow and the existing flows in the system.

We can find the sum of series that is included in Equation 4.20 in Table 4.3:

$$\sum_{i=1}^{m} \frac{2i-1}{2^{m-i}} = \frac{6}{2^m} + 4m - 6$$

Therefore, Equation 4.20 is reduced to:

$$l_q = \left(\left(\frac{t_a}{4} \cdot \left(\frac{6}{2^m} + 4m - 6 \right) \right) \cdot \left(1 - \frac{t_1}{2t_a} \right) + \left(\frac{2mt_a + t_1}{2} \right) \cdot \frac{t_1}{2t_a} \right) \cdot \xi \qquad (4.21)$$

Figure 4.14 illustrates, how the n_{sd} value changes, as time passes. A is the initial value for n_{sd}, namely ψ_{sd}^{o}, and B is the value that n_{sd} converges, namely ψ_{sd}. $n_{sd}(t_a)$ stands for n_{sd} value at time t_a.

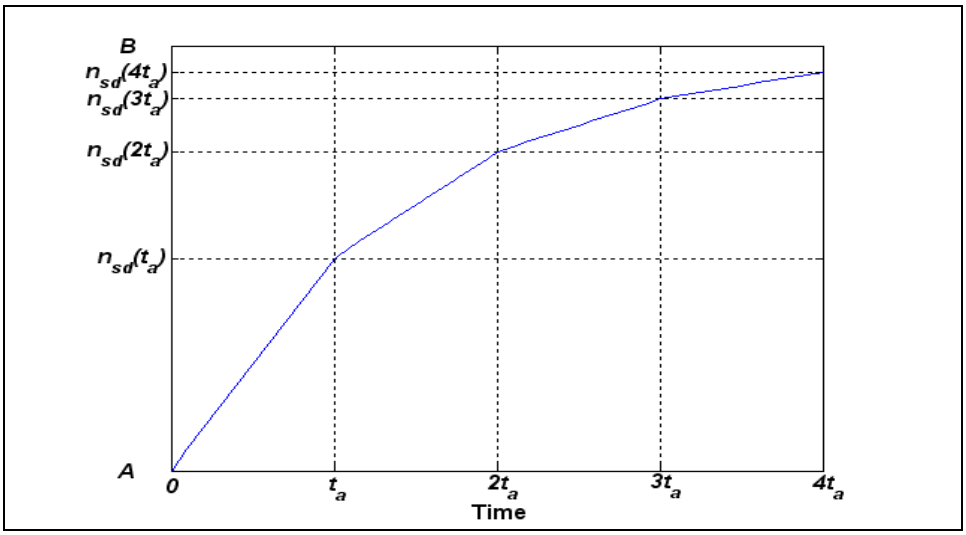


Figure 4.14. Illustration of Equation 4.19

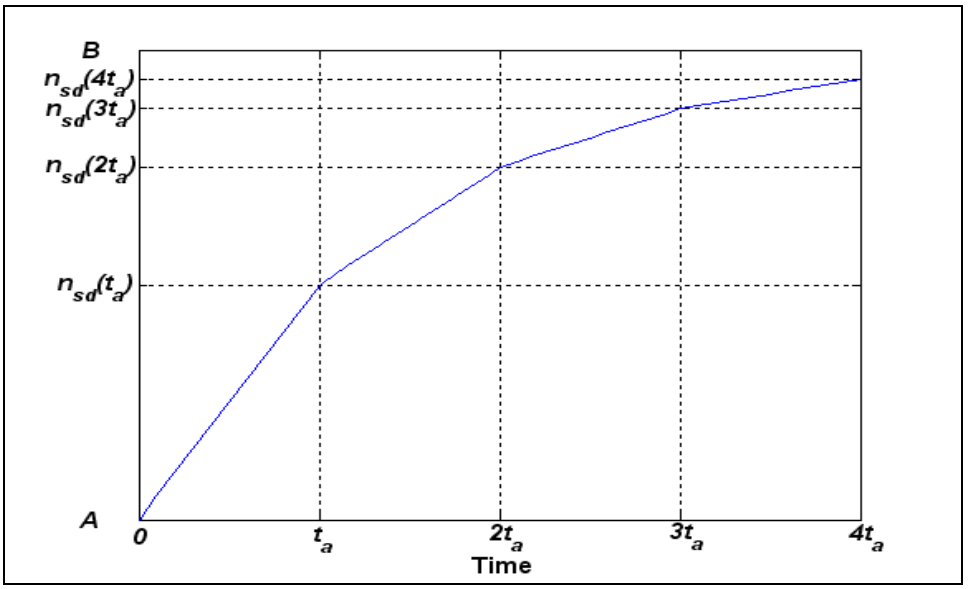


Figure 4.15. Illustration of Equation 4.21

Figure 4.15 illustrates the increment in the l_q value with time. $l_q\,(t_a)$ stands for the l_q value at time t_a. Even though the increment in the queue length is linear, the increment of l_q is not linear because of the effect of aging mechanism.

Figure 4.14 and Figure 4.15 illustrate the case that no other new flow is participated and none of the flows is redirected to another link. At some point, the corresponding link will become unfavored for the flows over that link, and redirections will occur. However, if all of the flows are redirected continuously, it can lead to congestion in alternative link too. In that case, all of them will again redirected to this link, and this will lead to needless fluctuations and disruption of stability, which will decrease the performance of the system. To avoid this scenario, we can redirect the flows with smaller data rates first, by adjusting the ePAR parameters. If l_q is too dominant for determining priority metric, we cannot achieve this. Difference between priority metrics for two flows f_{ij} and f_{kl} on the corresponding link is determined by:

$$\left|\mu_{ij} - \mu_{kl}\right| = \alpha \cdot \left|n_{ij} - n_{kl}\right| \tag{4.22}$$

The difference between n_{ij} and n_{kl} values is:

$$\left|n_{ij}-n_{kl}\right|=\left|\psi_{ij}^{\circ}-\psi_{kl}^{\circ}\right|\cdot\frac{1}{2^{m}}\cdot\left(1-\frac{t_1}{2t_{\mathrm{a}}}\right)+\left|\psi_{ij}-\psi_{kl}\right|\cdot\left(\frac{1}{2^{m}}+\frac{1}{2^{m}}\cdot\frac{t_1}{2t_{\mathrm{a}}}\right) \tag{4.23}$$

Considering that $\{m,n\}$ is the source-destination pair of the newly participating flow f_{mn}, for $\{i,j\}\neq\{m,n\}$ and $\{k,l\}\neq\{m,n\}$ case, Equation 4.23 is reduced to:

$$\left|n_{ij}-n_{kl}\right|=\left|\frac{r_{ij}-r_{kl}}{\lambda}\right|\cdot\frac{1}{2^{m}}\cdot\left(1-\frac{t_1}{2t_{\mathrm{a}}}\right)+\frac{B}{\lambda}\left|\frac{r_{ij}-r_{kl}}{X+r_{mn}}\right|\cdot\left(1-\frac{1}{2^{m}}+\frac{1}{2^{m}}\cdot\frac{t_1}{2t_{\mathrm{a}}}\right)$$

and it is equal to:

$$\left|n_{ij}-n_{kl}\right|=\left|\frac{r_{ij}-r_{kl}}{\lambda}\right|\cdot\left(\frac{1}{2^{m}}\cdot\left(1-\frac{t_1}{2t_{\mathrm{a}}}\right)+\frac{B}{X+r_{mn}}\cdot\left(1-\frac{1}{2^{m}}+\frac{1}{2^{m}}\cdot\frac{t_1}{2t_{\mathrm{a}}}\right)\right) \tag{4.24}$$

For sufficiently long aging periods, we can assume that redirections occur in first aging period, hence m is equal to zero. Then Equation 4.24 is reduced to:

$$\left|n_{ij}-n_{kl}\right|=\left|\frac{r_{ij}-r_{kl}}{\lambda}\right|\cdot\left(\left(1-\frac{t_1}{2t_{\mathrm{a}}}\right)+\frac{B}{X+r_{mn}}\cdot\frac{t_1}{2t_{\mathrm{a}}}\right) \tag{4.25}$$

For the case that $\{k,l\}=\{m,n\}$, Equation 4.23 reduces to following:

$$\left|n_{ij}-n_{mn}\right|=\left|\frac{r_{ij}}{\lambda}\right|\cdot\frac{1}{2^{m}}\cdot\left(1-\frac{t_1}{2t_{\mathrm{a}}}\right)+\frac{B}{\lambda}\left|\frac{r_{ij}-r_{mn}}{X+r_{mn}}\right|\cdot\left(1-\frac{1}{2^{m}}+\frac{1}{2^{m}}\cdot\frac{t_1}{2t_{\mathrm{a}}}\right) \tag{4.26}$$

For $m = 0$ case it is:

$$\left|n_{ij}-n_{mn}\right|=\left|\frac{r_{ij}}{\lambda}\right|\cdot\left(1-\frac{t_1}{2t_{\mathrm{a}}}\right)+\frac{B}{\lambda}\left|\frac{r_{ij}-r_{mn}}{X+r_{mn}}\right|\cdot\left(\frac{t_1}{2t_{\mathrm{a}}}\right) \tag{4.27}$$

As it was mentioned before, l_{q} should not be too dominant in determining priority metric. This suggests that difference occur between priority metrics for two flows, with

reasonably distinct data rates, should be near to the change occur in $\beta \cdot l_q$ value in a given reasonable time t_d.

Again assuming that $m = 0$, and initial l_q value is zero, difference occured in the l_q value in time t_d is the following:

$$l_q(t_d) - l_q(0) = \frac{t_d^2}{4t_a} \cdot \frac{X + r_{mn} - B}{\lambda} \tag{4.28}$$

Now, we can set:

$$\alpha \cdot \left| n_{ij}(t_d) - n_{kl}(t_d) \right| = \beta \cdot \left(l_q(t_d) - l_q(0) \right) \tag{4.29}$$

where f_{ij} and f_{kl} has reasonably distinct data rates, and t_d is chosen appropriately. If t_d is too small, then system suffers from concurrent redirections. In the other case, if it is too large, then the effect of queue length will be decreased in determination of the priority metric, and this can decrease the performance of the system. If the data rate of F_e is much higher than the data rate of f_{mn}, we can ignore the case that $\{k,l\}=\{m,n\}$, and thus Equation 4.29 can be rewritten as:

$$\frac{\alpha}{\beta} = \frac{\frac{t_d^2}{4t_a} \cdot \frac{X + r_{mn} - B}{\lambda}}{\left| \frac{r_{ij} - r_{kl}}{\lambda} \right| \left(\left(1 - \frac{t_d}{2t_a} \right) + \frac{B}{X + r_{mn}} \cdot \frac{t_d}{2t_a} \right)} \tag{4.30}$$

which reduces to:

$$\frac{\alpha}{\beta} = \frac{1}{2} \cdot \frac{t_d^2 \cdot \left(X + r_{mn} - B \right)}{\left| r_{ij} - r_{kl} \right| \left(\left(2t_a - t_d \right) + \frac{B}{X + r_{mn}} \right)} \tag{4.31}$$

If we set the reasonable rate difference $\left| r_{ij} - r_{kl} \right|$ to 0.5, Equation 4.31 reduces to:

$$\frac{\alpha}{\beta} = \frac{t_{\mathrm{d}}^2 \cdot (X + r_{mn} - B)}{\left((2t_{\mathrm{a}} - t_{\mathrm{d}}) + \frac{B}{X + r_{mn}} \right)} \tag{4.32}$$

Figure 4.16 illustrates the reasonable α/β values for different *X* values. Note that the *y* axis is equal to zero in the case that the total flow does not exceed the link capacity after the participation of new flow, and thus queue length does not increase ($X = B - r_{mn}$). In the case that $X = B$, the link was already fully utilized before the participation of new flow. At that point, corresponding α/β is equal to *R*. From Equation 4.32, the value of *R* is found to be:

$$R = \frac{t_{\mathrm{d}}^2 \cdot r_{mn}}{\left((2t_{\mathrm{a}} - t_{\mathrm{d}}) + \frac{B}{B + r_{mn}} \right)} \tag{4.33}$$

If the flow f_{mn} is not partitioned to various links, the expected value for r_{mn} is 1, and hence we may reduce Equation 4.33 to the following:

$$R = \frac{t_{\mathrm{d}}^2}{\left((2t_{\mathrm{a}} - t_{\mathrm{d}}) + \frac{B}{B + 1} \right)} \tag{4.34}$$

$X = B - r_{mn}$ and $X = B$ are two extreme points for *X*, and a reasonable *X* value should be chosen between these two values. For example, setting it to the mean value between $B - r_{mn}$ and *B*, that is $B - \frac{r_{mn}}{2}$ (or $B - 0.5$ if we set $r_{mn} = 1$), makes sense. The corresponding α/β value will indicate possible selection of ePAR parameters for optimal performance. However, for more accurate decision, actual traffic dynamics of particular networking scenario should be taken into account.

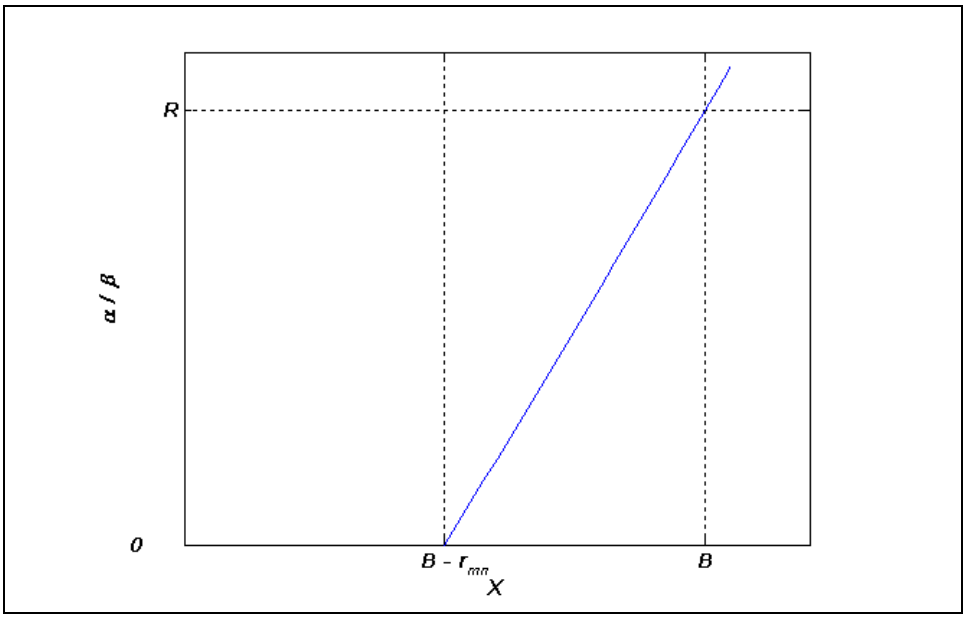


Figure 4.16. Illustration of Equation 4.32

Although the analysis given above may seem to be restricted with large buffer size and long aging duration, it can be easily enhanced and adapted for a wider domain. Therefore, tailoring this analysis for actual satellite constellations may be a worthy experiment for demonstrating its great benefits.

4.5. Summary

In this section, we first propose two traffic-sensitive shortest path routing algorithms for NGEO satellite networks, namely Priority-based Adaptive Routing (PAR) and enhanced PAR (ePAR). In the first algorithm (PAR), rather than setting the route in the terrestrial nodes or in a single satellite node, the route is set-up by making the decision of sending packet from which outgoing link at each hop. The decision criterion is based on a priority mechanism, which favors links that are relatively less utilized. The second algorithm (ePAR) is proposed to further enhance the routing algorithm for providing channeling of packets with same source-destination pairs to same links. The rationale behind this enhancement is that less contention may occur between packets over same routes. As part of the proposed algorithms, we introduce a new priority metric that includes some design parameters. To achieve a higher system performance, we present a detailed analysis of ePAR for adjusting the design parameters. Relying on extensive sets of

simulation results, we show that the proposed algorithms not only increase the throughput and link utilization, but also decrease the delay. Moreover, the proposed priority-based algorithms have no signaling overhead, and are promising for use in NGEO satellite networks.

In order to further increase the performance of the system, we propose a deflection routing mechanism, which deflects the packets to longer routes when the outgoing links in shortest paths are not available. Simulation results show that proposed deflection routing approach is promising for low traffic loads, but it fails to improve performance for high traffic loads. Including traffic load sensitivity to the deflection mechanism would be an interesting subject of a future study. Moreover, our deflection mechanism could be slightly modified for handling satellite failures.

5. NETWORK MOBILITY MANAGEMENT IN EARTH-FIXED NGEO SATELLITE SYSTEMS

5.1. Motivation and Related Work

Satellite networks are an attractive option to provide broadband integrated Internet services to globally scattered users, due to their potential advantages such as extensive geographic coverage, high bandwidth availability, and inherent broadcast capabilities. Satellites rotating in GEO are well suited for broadcast services; however, they suffer from high free space attenuation and long delays. On the contrary, non-geostationary systems consisting of MEO and LEO satellites offer lower latency, lower free space loss, and better re-use of available ground-space communication frequencies. Therefore, they are more suitable for most applications, especially, for those running in real-time. However, these advantages come with a price: Footprints of satellites at lower altitudes are smaller, and global coverage can be provided by higher number of satellites connected with ISLs. Moreover, lower orbit satellites move with higher speeds relative to the Earth's surface, resulting in high dynamic in the network topology. This topological phenomenon constitutes a major challenge in providing QoS for rapidly-growing real-time multimedia services.

Connectionless protocols may use the network resources efficiently; however, providing QoS guarantees is difficult in connectionless networks, in particular due to the difficulty in accounting for the delay aspects of QoS and sequencing. Reliable and powerful traffic engineering methods and QoS provisioning mechanisms are usually provided through connection orientation. However, connection-oriented protocols face an important challenge in satellite networks: Established connections must be maintained as the network topology changes. To address this challenge, satellite topology dynamics should be handled properly. Fortunately, although the topology of a satellite network rapidly changes, these changes are periodic and predictable because of the strict orbital movements of the satellites. Moreover, satellite constellation topology has a regular and highly symmetric structure. Considering these properties, some techniques are proposed in

order to make up with the mobility of the network topology. Virtual Node (VN) concept is one of the most common approaches. In the VN technique, a fixed virtual topology consisting of VNs is superimposed over the physical topology in order to hide the mobility of satellites from the routing protocols. To the best of our knowledge, it is first described by Mauger and Rosenberg [24], and many researchers develop their routing protocol based on this approach. However it has some deficiencies, which come from the fact that VN concept necessitates one-to-one correspondence between physical satellites and virtual nodes and it could not be applied for systems where more than one satellite can serve for a single footprint area. To make up with this, we propose and model a Multi-state Virtual Network (MSVN) topology which enables more than one satellite to cover a single footprint area. Details of MSVN are given in Section 5.3 together with brief discussion about its usefulness. To clarify potential advantages of MSVN-based systems over VN-based systems, firstly we focus on handover mechanisms for both kinds of systems. In Section 5.4, we first propose an efficient handover algorithm for VN-based systems. Next, we devise soft and semi-soft handover algorithms for MSVN-based systems. Comparison of proposed algorithms shows potential advantage of MSVN over single state conventional VN architecture. To our knowledge, this is the first work that investigates link-layer handovers in Earth-fixed satellite systems. In Section 5.5, we describe another possible approach, namely optimal beam management, for increasing system availability in MSVN-based Earth-fixed satellite systems. Following this, Section 5.6 concludes the chapter.

5.2. Satellite Network Dynamics

Low orbit satellites move with higher speeds relative to the Earth's surface, resulting in high variation in the satellite constellation topology. Therefore, connections between a terminal and a satellite must be handed over to another satellite when the current satellite drops too low above the horizon. On the other hand, handovers of active communications should be controlled considering the capabilities of the satellite antenna system. There are two general techniques:

- ***Asynchronous Handover***: It is generally appropriate for satellite antenna systems that have *satellite-fixed (nadir pointing) footprint*. As the satellite moves across the sky, its

footprint sweeps across the surface with a constant velocity, as described in Section 1.1.3 (see Figure 1.1(a)). When a terminal reaches the edge of the current footprint, it is handed off to a new satellite whose footprint is entering the area.

- **S*ynchronous Handover*:** This approach can be applied for the satellite systems where a satellite is capable of electronically steering its beams, so that it can make up for its motion and the satellite footprint can be fixed for a time duration. This leads to *Earth-fixed footprints* as described in Section 1.1.3 (see Figure 1.1(b)). After some time, all the satellites will be moving away from their corresponding footprints, and the system periodically reassigns each satellite to a new fixed footprint. This approach is proposed in [9], and it is called synchronous handover since all handovers occur simultaneously. Although this technique comes with the cost of degradation in the elevation mask used in the system (or increase in the number of satellites), it has the potential to simplify the handovers. In this chapter, we consider satellite systems with Earth-fixed footprints.

Synchronous handover technique eases the mobility management issue, since it allows each physical satellite to be represented by a VN. In the VN technique, a fixed virtual topology consisting of VNs is superimposed over the physical topology in order to hide the mobility of satellites from the network layer. A logical address is assigned to the fixed portions of the Earth's surface. Then, by using *Earth-fixed* satellite systems described above, a satellite embodies the VN above this fixed Earth footprint for the time period during which it is serving that footprint. Each VN is embodied at any given time by a certain physical satellite. As a satellite disappears over the horizon, its corresponding VN becomes represented by the next satellite passing overhead and the state information (such as routing table entries or channel allocation information) is transferred to it. Handover between VNs and physical satellites are synchronously performed, hence, the virtual topology remains unchanged. A routing decision is made on this fixed virtual topology, and consequently, the network layer is isolated from the satellite constellation dynamics.

Recently, many routing protocols are proposed based on VN concept. A mechanism to adopt IP routing at the VNs in order to seamlessly integrate space network with terrestrial Internet and provide direct support for IP-QoS and IP-Multicast is presented in

[25]. A distributed datagram routing algorithm, and a multicast routing algorithm regarding satellite network as a mesh topology consisting of fixed logical locations (virtual nodes) are introduced in [26], and [27], respectively. Moreover, some hierarchical routing algorithms that are developed for integrated satellite networks consisting of LEOs and MEOs [28] or HAPs, LEOs and GEOs [30] simplify LEO layer by modeling it as a fixed virtual network.

Although VN concept is widely accepted, it has also some potential drawbacks that need to be significantly improved. Representing each physical satellite with a virtual node implies one-to-one mapping between terminals in a given footprint and satellite serving to that footprint. Conventional VN concept does not allow multiple satellites to serve single footprint area. However, this is desirable property due to following reasons:

1. A single satellite may not be sufficient to serve all terminals in a particular footprint area due to high service demand of ground terminals (GTs) or because of shadowing by terrain and buildings. Therefore, system availability can be increased by providing coverage to an area through multiple satellites.

2. During daylight hours, if the satellite is located along the same line of sight with the Sun, communication becomes impossible due to the Sun's radiation overwhelming the satellite signal. This is called sun outage, and occurs around the time of the spring and fall equinoxes when the Sun crosses the Earth's equatorial plane. System can compensate these situations only by offering alternative satellites.

3. More bandwidth can be provided for densely populated areas by directing beams of neighboring satellites to these areas.

4. To achieve lossless handover, at least two satellites should serve a terminal during handover task. VN technique eliminates the need for network layer handover but in lower layers, it could not support soft handover, and only solution is a lossy hard handover.

Considering these issues, the necessity of one-to-one correspondence between VNs and physical satellites constitutes major shortcomings of conventional VN approach. Therefore, we propose the *multi-state virtual network (MSVN)* concept that enables more than one satellite to cover a single footprint area. We describe this concept in next section.

5.3. Virtual Topology Dynamics

In our analysis of virtual topology dynamics, we consider a satellite network with N_P orbit planes, and each orbit consists of N_{SAT} satellites that have Earth-fixed footprints, and serve for total of N_{FP} footprint areas along the orbit. We also assume that inter-satellite distances are equal for the satellites in same orbit. In conventional VN concept, N_{SAT} is equal to N_{FP}, i.e., there exists a single satellite for each footprint. However, for the sake of increasing system availability, we also consider the cases where there is more than one satellite per footprint.

Definition 5.1 (*Average number of satellites per footprint*): Average number of satellites per footprint area ($N_{S/F}^{avg}$) is calculated as:

$$N_{S/F}^{avg} = \frac{N_{SAT}}{N_{FP}} \tag{5.1}$$

For conventional VN concept, N_{SAT} is equal to N_{FP} and $N_{S/F}^{avg}$ is equal to 1.

The case where $N_{S/F}^{avg} = 2$ is illustrated in Figure 5.1(a). In this case, as a satellite leaves the footprint area, another one starts serving the same area. Therefore, each footprint is always served by two satellites. Hence, we can model the entire system as a fixed virtual network as in Figure 5.1(b)[1]. A circle denotes a virtual node that is embodied by a single physical satellite, e.g., v_1 in Figure 5.1(b) is embodied by satellite 1 in Figure 5.1(a).

[1] Note that we assume that satellites in different orbits switch footprint areas synchronously.

Rectangles represent footprint areas that include ground terminals (GT) served by two satellites (that are represented by two circles within the rectangular area) at a time. Note that physical satellites are connected with GTs by up-down links (UDLs); therefore, it can be considered that there are undesignated links between rectangle nodes and the circle nodes covered by them. In this case, system availability is improved compared to the conventional VN approach; however, satellites need to transfer the state information two times more frequently.

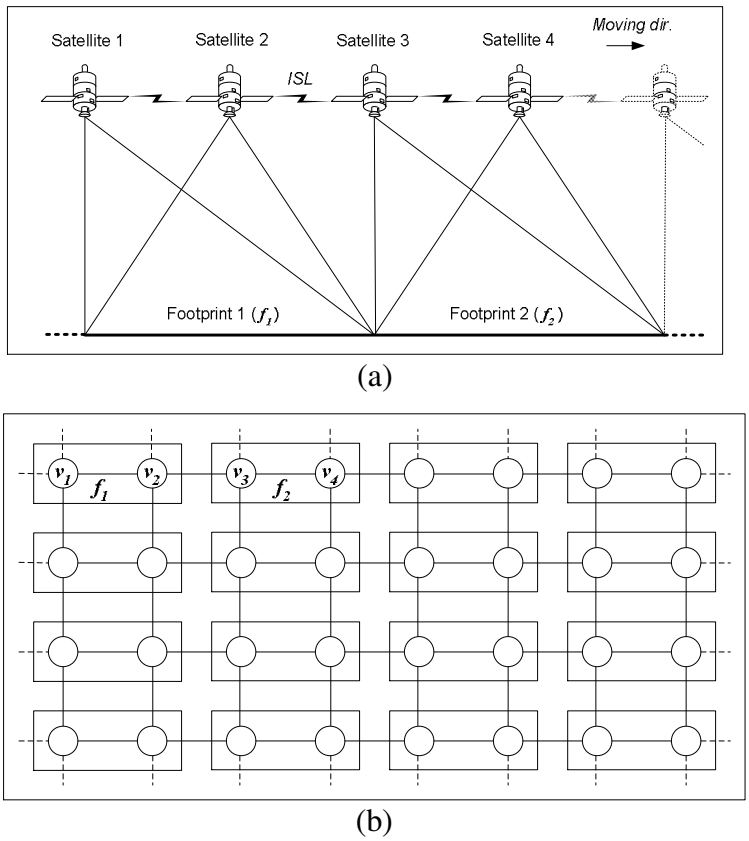
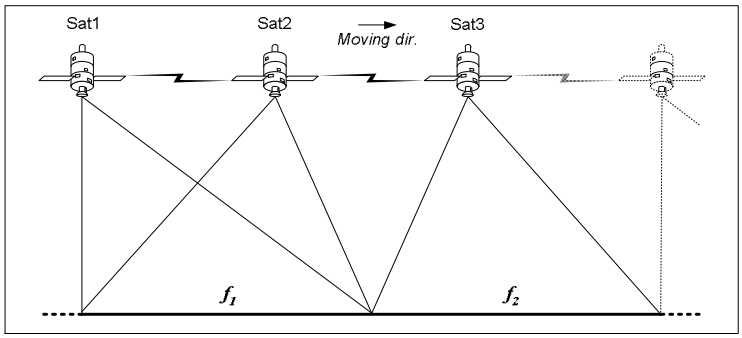


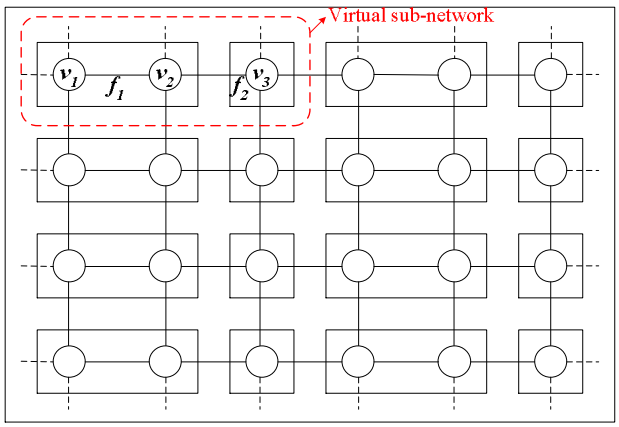
Figure 5.1. (a) Part of a satellite system with $N_{\mathrm{S/F}}^{\mathrm{avg}} = 2$ and (b) corresponding virtual network

Note that, when $N_{\mathrm{S/F}}^{\mathrm{avg}}$ is an integer value, then we have a fixed virtual topology and each footprint area is always served by $N_{\mathrm{S/F}}^{\mathrm{avg}}$ satellites. However, for the systems where $N_{\mathrm{S/F}}^{\mathrm{avg}}$ has a non-integer value, number of satellites covering a footprint area ($N_{\mathrm{S/F}}$) changes with time, which yields a dynamic virtual topology. In such systems, at a given time, a footprint area is served by $\lfloor N_{\mathrm{S/F}}^{\mathrm{avg}} \rfloor$ or $\lceil N_{\mathrm{S/F}}^{\mathrm{avg}} \rceil$ satellites. The case with $N_{\mathrm{S/F}}^{\mathrm{avg}} = 1.5$ where three

satellites serve for two footprints every time is illustrated in Figure 5.2(a). Two satellites (Sat1 and Sat2) serve f_1 and one (Sat3) serves f_2. Corresponding virtual network is VNet1 shown in Figure 5.2(b). (Virtual sub-network in the figure corresponds to the part of the network shown in Figure 5.2(a).) However, when Sat2 switches to f_2 as shown in Figure 5.3(a), then two satellites (Sat2 and Sat3) continue serving f_2 and only Sat1 serves f_1, and virtual topology changes to VNet2 shown in Figure 5.3(b). After a while, new satellite enters f_1 area, and Sat3 leaves f_2 and the virtual topology switches to the VNet1. In this case, virtual topology has two states, one is active and the other is passive at a given time. For example, in the situation shown in Figure 5.2(a), VNet1 is active and VNet2 is passive, and for Figure 5.3(a), VNet2 is active and VNet1 is passive. We call such virtual topology with multiple states, *Multi-state Virtual Network (MSVN)* topology. Number of states depends on $N_{\mathrm{S/F}}^{\mathrm{avg}}$ value. In the next section, we describe general case of MSVN and provide a formal model.

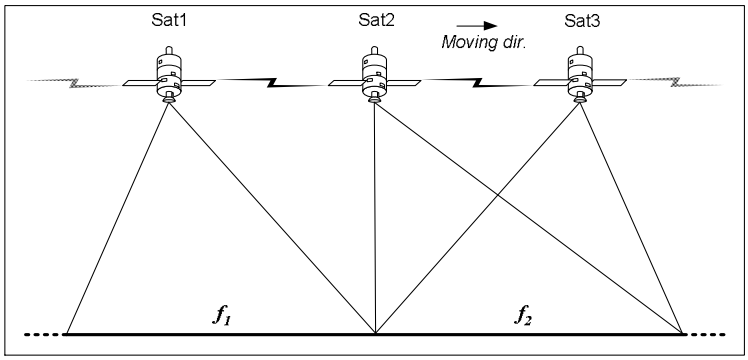


(a)

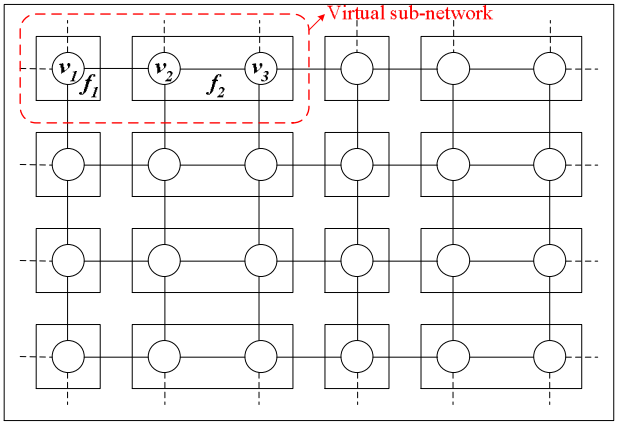


(b)

Figure 5.2. (a) Part of a satellite system with $N_{\mathrm{S/F}}^{\mathrm{avg}} = 1.5$: case 1 and (b) corresponding virtual network (VNet1)



(a)



(b)

Figure 5.3. (a) Part of a satellite system with $N_{\mathrm{S/F}}^{\mathrm{avg}} = 1.5$: case 2 and (b) corresponding virtual network (VNet2)

5.3.1. Multi-state Virtual Network (MSVN) Topology

In order to provide a formal description, we give set of definitions.

Definition 5.2 (Inter-satellite Angular Distance): Consider that N_{SAT} satellites exist per satellite orbit. If we assume that inter-satellite links between each satellite in the same orbit are identical and the angular length of an orbit is 2π, then, the angular distance (L_{SAT}) between each satellite pair is expressed by

$$L_{\mathrm{SAT}} = \frac{2\pi}{N_{\mathrm{SAT}}} \tag{5.2}$$

Definition 5.3 (Angular length of an Earth footprint): Let all of the satellites in an orbit serve for N_{FP} footprint areas. Considering that the sum of the angular lengths of these footprints is 2π, the angular length of a single footprint (L_{FP}) is calculated as

$$L_{FP} = \frac{2\pi}{N_{FP}} \tag{5.3}$$

Note that actually footprints are circular areas, and some overlapping between the footprints of the adjacent satellites is necessary. Therefore, we consider *effective footprint* [35] of a satellite which is equivalent to the largest hexagon inscribed into the footprint as shown in Figure 5.4.

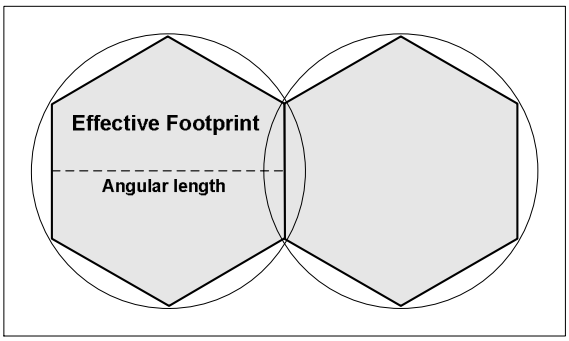


Figure 5.4. Effective footprint and its angular length

Definition 5.4 (Minimum area served by fixed number of satellites): In conventional VN concept, single footprint area is served by fixed number of (one) satellite all the time. In MSVN, minimum number of consecutive footprints served by fixed number of satellites (M_{FP}) changes with the value of $N_{S/F}^{avg}$. When $N_{S/F}^{avg}$ is integer, M_{FP} is equal to one. For $N_{S/F}^{avg} = 1.5$, M_{FP} is two as shown in Figure 5.2 (a). In general, M_{FP} can be found by

$$M_{FP} = \frac{\text{lcm}(\eta L_{SAT}, \eta L_{FP})}{\eta L_{FP}} \tag{5.4}$$

where lcm is the least-common-multiple function and η is a normalization factor, such that both ηL_{SAT} and ηL_{FP} are integer values. If we set

$$\eta = \frac{N_{SAT} \cdot N_{FP}}{2\pi} \tag{5.5}$$

then (5.4) reduces to

$$M_{FP} = \frac{\text{lcm}(N_{FP}, N_{SAT})}{2\pi} \tag{5.6}$$

M_{FP} consecutive footprint areas are always served by M_{SAT} satellites. Therefore, M_{SAT} can be expressed by

$$M_{SAT} = M_{FP} \cdot N_{S/F}^{avg} = \frac{\text{lcm}(N_{FP}, N_{SAT})}{N_{FP}} \tag{5.7}$$

Definition 5.5 (High and low service modes): Recall that the number of satellites serving a footprint area ($N_{S/F}$) varies between $N_{S/F}^{L}$ and $N_{S/F}^{H}$, where $N_{S/F}^{L} = \lfloor N_{S/F}^{avg} \rfloor$ and $N_{S/F}^{H} = \lceil N_{S/F}^{avg} \rceil$. A footprint area is said to be in *low service mode* when $N_{S/F}^{L}$ satellites serve it and in high service mode when $N_{S/F}^{H}$ satellites serve. At a given time, among M_{FP} consecutive footprints, M_{FP}^{L} footprints are served by $N_{S/F}^{L}$ satellites and M_{FP}^{H} footprints are served by $N_{S/F}^{H}$ satellites. When $N_{S/F}^{avg}$ is integer, M_{FP}^{L} and M_{FP}^{H} are equal to M_{FP} (which is one). Otherwise,

$$M_{FP}^{L} = M_{FP} \cdot \left(\lceil N_{S/F}^{avg} \rceil - N_{S/F}^{avg} \right) \tag{5.8}$$

$$M_{FP}^{H} = M_{FP} \cdot \left(N_{S/F}^{avg} - \lfloor N_{S/F}^{avg} \rfloor \right) \tag{5.9}$$

In order to validate that both M_{FP}^{L} and M_{FP}^{H} have integer values, recall that

$$\lfloor N_{S/F}^{avg} \rfloor = \left\lfloor \frac{N_{SAT}}{N_{FP}} \right\rfloor = \frac{N_{SAT} - (N_{SAT} \bmod N_{FP})}{N_{FP}} \tag{5.10}$$

Using (5.6), (5.9) and (5.10),

$$M_{\mathrm{FP}}^{\mathrm{H}} = \frac{\mathrm{lcm}\left(N_{\mathrm{FP}}, N_{\mathrm{SAT}}\right)}{N_{\mathrm{SAT}}} \cdot \frac{N_{\mathrm{SAT}} \bmod N_{\mathrm{FP}}}{N_{\mathrm{FP}}} = \frac{N_{\mathrm{SAT}} \bmod N_{\mathrm{FP}}}{\gcd\left(N_{\mathrm{SAT}}, N_{\mathrm{FP}}\right)} \tag{5.11}$$

which is known to be an integer value. In a similar manner, the following equation is derived for low service mode:

$$M_{\mathrm{FP}}^{\mathrm{L}} = \frac{N_{\mathrm{FP}} - N_{\mathrm{SAT}} \bmod N_{\mathrm{FP}}}{\gcd\left(N_{\mathrm{SAT}}, N_{\mathrm{FP}}\right)} \tag{5.12}$$

which is also clearly an integer value.

Definition 5.6 (Multi-state virtual sub-network): An MSVN consists of identical multi-state virtual sub-networks (MSVSN) connected with each other. An MSVSN consists of M_{FP} consecutive footprint areas (served by M_{SAT} satellites), M_{SAT} virtual nodes (each embodied by a satellite) and $M_{\mathrm{SAT}} - 1$ links (which connects consecutive virtual node pairs). Note that (6) suggests N_{FP} to be an integer multiple of M_{FP}, i.e.,

$$N_{\mathrm{FP}} = r \cdot M_{\mathrm{FP}},\ r \in Z^{+} \tag{5.13}$$

Each of these r consecutive footprint groups are served by M_{SAT} satellites in an identical manner. Therefore, for an orbit, there exists r identical MSVSNs. All of these sub-networks have the same number of states that switch synchronously. Moreover, recall that we assume that satellites in different orbits switch their footprint areas in a synchronous manner. Therefore, if the satellite network has N_{P} orbit planes, then an MSVN consists of $N_{\mathrm{P}} \cdot r$ identical MSVSNs. Hence, modeling single MSVSN and linking $N_{\mathrm{P}} \cdot r$ such sub-networks are sufficient to model an MSVN. Links between sub-networks are determined by the ISL characteristics of physical network.

Definition 5.7 (System period and state intervals): System period (T_S) is defined as the time elapsed for a virtual topology to repeat its states. For a system with only one state, T_S is equal to ∞. For systems with more than one state, T_S is the following:

$$T_S = \Delta t_{ST} \cdot N_{ST} \tag{5.14}$$

where Δt_{ST} is the state interval, i.e., the time period for changing a state. N_{ST} is the number of states, which is shown to be equal to M_{FP} with the theorem given next.

Theorem 5.1: For an MSVN, N_{ST} is equal to M_{FP}.

Proof: As mentioned before, the number of states for an MSVN is equal to the number of states for its MSVSN. Consider a moment when a satellite just starts to serve the first footprint area of MSVSN, and suppose that the system is in *state 1* at that moment. This state is repeated when the next following satellite starts to serve this first footprint. Between these two events, system changes its state N_{ST} times. Note that the state changes only when a satellite switches its footprint. In an MSVSN, two satellites cannot switch to a new footprint at the same time (otherwise there would be two consecutive footprint groups of size $M_{FP}/2$ served by fixed number of satellites). Moreover, each of the points over an orbit is passed by exactly one satellite during the system period. Therefore, exactly one satellite switches to each of the M_{FP} footprints. Since there are M_{FP} consecutive footprints and exactly one satellite switches to each of these footprints (in different times) during a system period, there are exactly M_{FP} states in the system.

The number of states is illustrated in Figure 5.5 for various N_{SAT} values when N_{FP} is equal to 24.

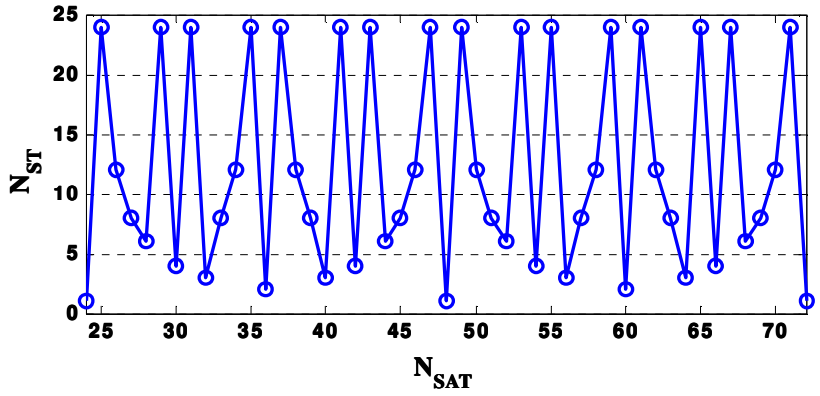


Figure 5.5. Number of states versus N_{SAT} (N_{FP} = 24)

At a given time, only one state is active and others are passive. The active state changes at each Δt_{ST} time units, and Δt_{ST} can be found as

$$\Delta t_{ST} = \frac{2\pi}{\text{lcm}(N_{FP}, N_{SAT}) \cdot \omega_{SAT}} \tag{5.15}$$

where ω_{SAT} stands for angular velocity of a satellite, and it is same for all satellites in the network.

Definitions given above clearly identify characteristics of an MSVN topology. To briefly summarize, MSVN consists of $N_P \cdot r$ identical MSVSNs. Each MSVSN can be considered as a set of virtual subnetworks (each corresponds to a particular state and remains active for Δt_{ST} time units). Each virtual subnetwork consists of M_{SAT} VNs, M_{FP} footprints (which can be considered as terrestrial nodes that are connected with VNs), set of links between neighboring VNs and set of links between VNs and terrestrial nodes (footprints). Footprints in high service mode are linked to more VNs than footprints in low service mode. At any given time, M_{FP}^{H} of the M_{FP} footprints are in high service mode, and M_{FP}^{L} are in low service mode. Service modes of footprint areas may change at each state in a deterministic manner.

A detailed formal mathematical model for MSVN and MSVSN is available in [14]. For the sake of clarity, we will not include further details in this thesis. Rather, we will discuss usefulness of MSVN topology and show some of its potential advantages.

5.3.2. Discussion

For the cases where $N_{S/F}^{avg}$ is not an integer value, a GT will be served by variable number of satellites. When a particular footprint area switches from high service mode to low service mode, system availability decreases, and this may force some of the connections to be dropped. Therefore, a call admission mechanism need to be employed in order to guarantee connectivity when the system changes its state. This means that in the

high service mode UDL links could not be fully utilized. Therefore, it could be argued that a satellite system (say *system 1*) with $N_{S/F}^{avg} = \gamma$ (where $\gamma > 1$ and $\gamma \notin Z$) has no further advantage compared to another system (say *system 2*) with $N_{S/F}^{avg} = \lfloor \gamma \rfloor$. Nevertheless, system 1 can offer higher system availability than system 2 due to several reasons elaborated next.

In system 1, effect of shadowing by terrain and buildings and sun outage problem are lower. This is because elevation angle between terminal and the nearest satellite is always greater in system 1. Moreover, on the average, there is more opportunity in selecting a satellite terminal to send data.

By letting some non-critical communications to perform only in high service mode, system utilization improves.

By properly adjusting beam directions, a single state virtual network topology can be achieved for system 1. In other words, by making satellites to direct their beams to neighboring footprint areas (with higher traffic density) we can satisfy the condition that every footprint area continuously served by fixed number of satellites. This significantly increases the system availability. We justify this idea in Section 5.5, where we propose a traffic aware optimal beam management technique, which significantly increases availability and utilization of the system with a marginal increase in the cost.

Moreover, it is possible to provide faster and smoother handover algorithms in system 1. As an example, soft handover could be employed in system 1, whereas it could not be employed in system 2, because all handovers occur simultaneously. In the next section, we focus on this issue, and investigate, propose and compare possible handover mechanisms for VN-based systems and MSVN-based systems.

5.4. Handover Mechanisms

As mentioned in Section 5.2, Earth-fixed satellite systems can offer synchronous handover, which means all handovers between satellites and fixed Earth terminals occur

simultaneously. Using VN technique, network layer handover could be totally eliminated because VNs have fixed IP addresses and routing mechanisms are utilized over the fixed virtual topology. However, although there is no need for network layer handover, physical network is dynamic and handover in lower layers could result in significant packet loss. Designing a smooth (low packet loss) and fast (low latency) handover algorithm is a crucial issue. Especially for satellite environments with long propagation delays, system performance could be significantly improved by using proper handover mechanisms. Handovers in satellite systems with satellite-fixed footprints are widely investigated in the literature, but, to our knowledge, there is no previous work that investigates link layer handovers in Earth-fixed satellite systems [59]. Therefore, we first propose an efficient handover scheme for VN-based satellite systems, namely Virtual Node Handover (VN-HO) algorithm. Next, we propose soft handover algorithm and semi-soft handover algorithm for MSVN-based satellite networks; MSVN-SHO and MSVN-SSHO. Comparison of proposed algorithms shows possible advantage of MSVN over single state conventional VN architecture.

5.4.1. Handover Mechanisms in VN-based Satellite System

In the VN technique, mapping between VNs and physical satellites are changed periodically. When a satellite changes its corresponding VN, it transfers state information, such as routing table entries and channel allocation information to the next satellite passing overhead and receives new state information from the previous satellite that embodies its new VN. In order to reduce latency, each satellite activates new state information simultaneously. However, as satellites change their state and address, the packets on the network will not be able to reach their intended target. To handle this problem, one simple method is to ignore these packets and retransmit them after new mapping between VNs and satellites is established. However, this method results in high handover latency and significant performance decrease. Instead, we propose a new handover algorithm for VN-based Earth-fixed satellite systems, namely Virtual Node Hand-Over (VN-HO) algorithm.

5.4.1.1. Virtual Node Handover (VN-HO) Algorithm. As we mentioned above, VN-based systems do not need network layer handover. However, link layer handover can take

significant amount of time. As a satellite switches from one footprint area to another, it should activate its new state information, steer its beams to serve for the new area, and start communicating to GTs in that area. VN-HO algorithm aims to minimize packet loss during this switching period. We can describe VN-HO algorithm as follows.

After receiving new state info from their predecessors, all satellites activate new state info in a simultaneous manner. As soon as new state information is activated, a satellite sends ACTIVATED message to all its neighbors to inform that packets sent after this time are according to new state. At the same time, it sends a STOP message to the GTs together with the information about the last packets that were successfully received. Afterwards, satellite stops receiving packets from and sending packets to GTs and starts steering its beams to new footprint area. Time needed to switch from one footprint area to other is called *switching time* (T_{SW}). During switching time, it is not possible to communicate with GTs. When a GT receives a STOP message it stops sending data and waits for the successor satellite to be ready.

In the new state, each satellite receives an ACTIVATED message from its neighbors and is informed that packets that are received after this message should be handled according to new state information. However, some of the packets are expected to be received before ACTIVATED message. These packets are checked and if the next hop is its new address, then they are forwarded according to the new routing table. Otherwise, packets are directed to successor satellite. Packets received after ACTIVATED message are forwarded according to the new routing table. If the next hop is GT but the switching time has not ended yet, packets are buffered. As soon as switching is performed and the relevant READY message is transmitted, these packets are forwarded to GTs. When GT receives a READY message, communication continues between new satellite and GT.

The proposed handover mechanism is illustrated in Figure 5.6 and Table 5.1. New satellite starts serving the footprint area where GT resides, and the GT hands off from old satellite to new satellite.

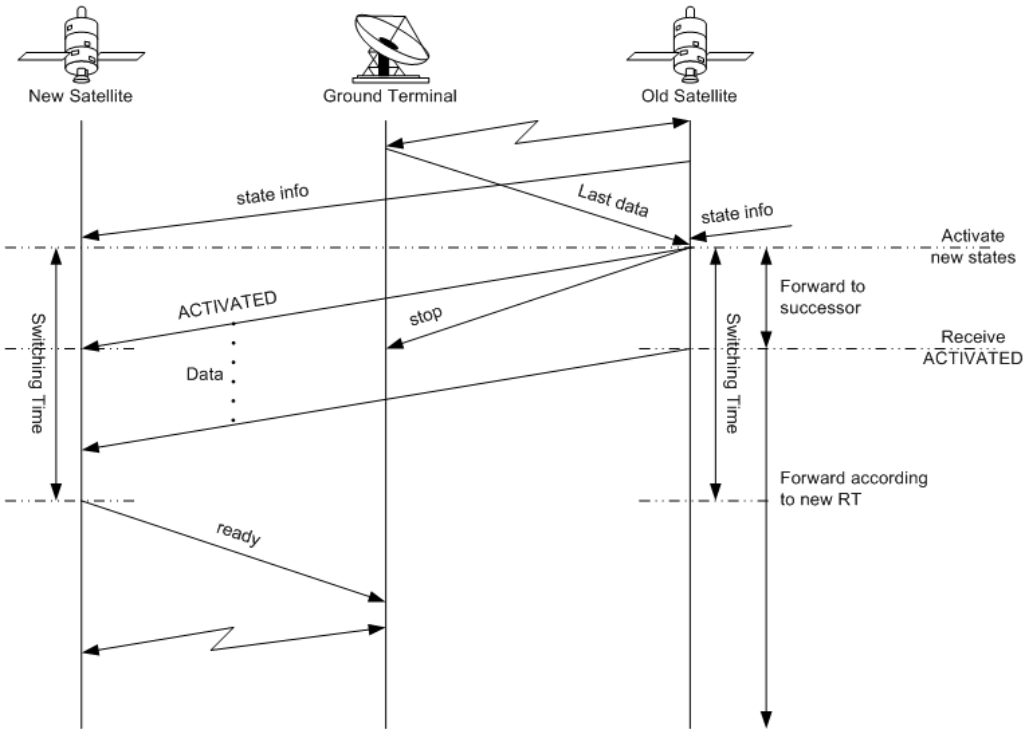


Figure 5.6. Illustration of VN-HO data and message flows

Table 5.1. VN-HO operation

Inter-Satellite signal and data flow
Step 1. Send state info to successor.
Step 2. Receive state info from predecessor and activate new state.
Step 3. Send ACTIVATED to neighbors.
For each incoming interface:
Step 4. Forward arriving packets to successor until receiving ACTIVATED*.
Step 5. Receive ACTIVATED and forward according to new routing table.
**Step 4 has an exception. If the next hop for the packet is the new address of the satellite, it is forwarded according to the new routing table.*
Satellite-GT signal and data flow
Step 1. Activate new state and send STOP to GTs.
-- Communication stops during switching time
Step 2. At the end of switching time, send READY to new corresponding GTs.
-- Communication continues…

Now, we analyze the handover latency of the proposed mechanism. Consider a flow that originates from a ground terminal G1, and sent to another ground terminal G2 via a satellite constellation. End-to-end handover latency ($L_{E\text{-}E}$) experienced by such a flow is

determined by latency in uplink UDL (L_{UDL_U}), latency in inter-satellite links (L_{ISL}), and latency in downlink UDL (L_{UDL_D}). For the VN-HO algorithm, amount of these latencies are as following:

$$L_{UDL\_U} = 2 \cdot T_{UDL} + T_{SW} \tag{5.16}$$

$$L_{UDL\_D} = \begin{cases} T_{ISL} & \text{if } T_{ISL} > T_{SW} \\ T_{SW} & \text{otherwise} \end{cases} \tag{5.17}$$

$$L_{ISL} = T_{ISL} \tag{5.18}$$

where T_{UDL} is UDL link delay, and T_{ISL} is intra-orbit ISL link delay. We assume that T_{UDL} and T_{ISL} values do not vary for different satellite-GT or satellite-satellite pairs.

$L_{E\text{-}E}$ is the maximum of these three latency values, i.e.,

$$L_{E-E} = \begin{cases} T_{ISL} & \text{if } T_{ISL} > 2 \cdot T_{UDL} + T_{SW} \\ 2 \cdot T_{UDL} + T_{SW} & \text{otherwise} \end{cases} \tag{5.19}$$

In most satellite systems, the first condition in Equation 5.19 is very unlikely to hold, therefore 5.19 reduces to

$$L_{E-E} = 2 \cdot T_{UDL} + T_{SW} \tag{5.20}$$

VN-HO eliminates losses inside the satellite network and only packet loss occurs in the UDLs. Packet loss in the uplink UDL occurs due to the fact that GT sends data to the satellite without caring about when the satellite will cut off the communication. Since satellite notifies the last packet that it received, recovering the packet losses is a relatively easy task. To avoid useless transmissions, satellite may send STOP message to GTs, at an appropriate time before breaking the link.

Packet loss in the downlink, may occur only when $T_{ISL} < T_{SW}$. In this case, VN-HO eliminates packet loss by buffering packets until the end of switching time. If T_{SW} is too long such that all packets could not be buffered, then some of the data packets are lost.

Receiver could receive some of the packets out of order. This is due to the fact that number of hops for a flow can vary because of the changes in the mapping between VNs and physical satellites.

5.4.2. Handover Mechanisms in MSVN-based Satellite System

In a VN-based system with $N_{S/F}^{avg} = 1$, GTs cannot be served during switching period. However, for the satellite systems where $N_{S/F}^{avg} > 1$, handover could be done in a smoother way. Without loss of generality, let us assume a satellite system where N_{SAT}=10, and N_{FP}=8. In such a network, four footprints are served by five satellites at any time, and the corresponding MSVSN topology for four states is shown in Figure 5.7.

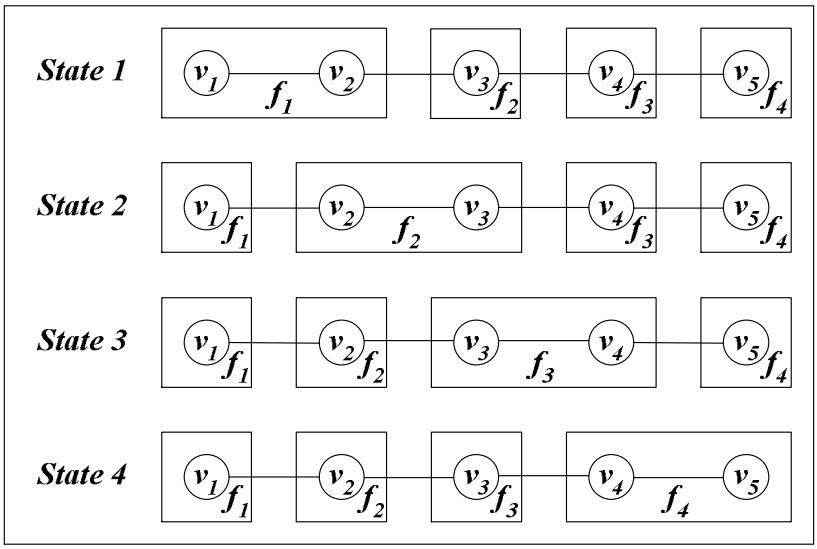


Figure 5.7. MSVSN topology for four states of a system with N_{SAT}=10 and N_{FP}=8

At the beginning, two satellites serve for f_1. Then one of them switches to f_2 and two satellites continue serving f_2. Afterwards, one of those switches to f_3, and so on. Each footprint area is served by two satellites for Δt_{ST} amount of time, and by one satellite for

$3\Delta t_{ST}$ amount of time. As described in Section 5.3.2, this condition can be used to facilitate soft handover. For this purpose, we investigate possible handover mechanisms for MSVN-based satellite networks.

In soft handover, GT is connected to the old satellite during the time for setting up connection with the new satellite, i.e., it is temporarily connected to both the old satellite and the new satellite. Afterwards, connection with old satellite is broken and communication continues with new satellite. In order to support soft handover, a GT should be capable of transmitting signals to and receiving signals from more than one satellite at a time (as in CDMA systems). If this is not the case, alternative solutions should be considered to enable a smooth handover scheme.

Here, we introduce two new handover algorithms: Soft handover algorithm for MSVN-based satellite networks (MSVN-SHO) works for systems where GTs are capable of communicating with more than one repeater at a time. Second algorithm is semi-soft handover algorithm (MSVN-SSHO) which works for systems where GTs have no such capability.

5.4.2.1. MSVN-SHO Algorithm. Assume a satellite system with non-integer $N_{S/F}^{avg}$ value, where $1< N_{S/F}^{avg} <2$. In such a system, a footprint area is served by one or two satellites at any time. Since UDL links could not be fully utilized in high service mode, as described in Section 5.3.2, we consider taking advantage of high service mode for achieving soft handover, rather than increasing transmission rate. By default, one of the satellites is connected to GTs and during handover both establish a connection temporarily. Then, new satellite continues serving GTs and the old satellite breaks off the UDL links. However, if needed, both satellites may serve simultaneously to increase system availability.

We define two modes for satellites: *Active* and *passive*. In the active mode, satellites communicate with GTs and forward packets according to the routing tables. In passive mode, satellites do not communicate with GTs and act as nodes that relay packets to active satellites. When two satellites (one active and one passive) passes over the same footprint

area, passive one is called the *passive counterpart* of the active satellite, and active one is called the *active counterpart* of the passive satellite.

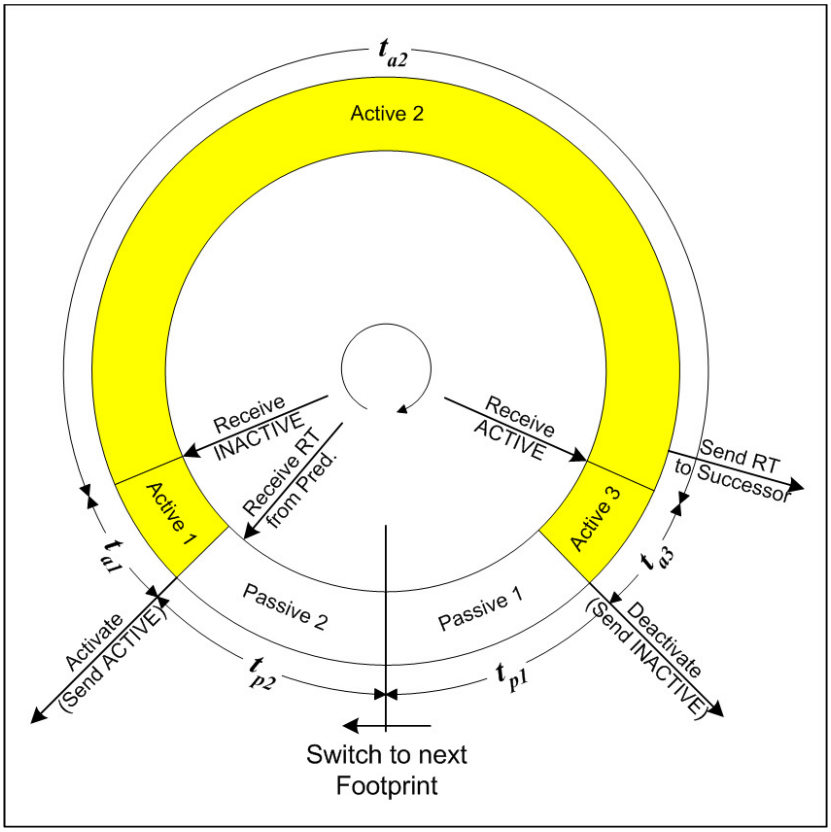


Figure 5.8. MSVN-SHO state diagram for a satellite

Now, we will describe MSVN-SHO operation based on these definitions. For each satellite, three active and two passive states are defined. The state diagram for a satellite is illustrated in Figure 5.8. As a satellite steers its beams to a new footprint area, it is in passive state. In this state (*passive 1*), the satellite does not have any active routing table and packets are just relayed to active satellites as stated in Table 5.2. Note that interface names are defined in Figure 5.9. Subsequently, the satellite receives the routing table and other state information (such as channel allocation info, etc) from its predecessor. After updating tables, it switches to the active mode and informs its predecessor and GTs by sending an ACTIVE message. In the active mode, the satellite forwards all the packets according to the routing table. Active duration is considered as three parts as depicted in Figure 5.8. When the predecessor satellite enters the passive mode, it sends an INACTIVE message. Since the packets that are received before this INACTIVE message are already

forwarded according to the routing table, it is not necessary to look up the routing table again and these packets can be just relayed without changing their direction (they arrive from interface *b* and forwarded to interface *a*). After INACTIVE message is received, the satellite switches to *active 2* state and forwards all the packets according to the routing table. The satellite switches from active 2 to active 3 state when it receives an ACTIVE message from the successor. This message informs that the successor is an active counterpart henceforth, and the packets that are received from interface *a* can just be relayed to interface *b* without referring to the routing table. After a while, the satellite switches to passive mode by deactivating the UDL links and sends an INACTIVE message to the successor. At this state (passive 2), packets received through intra-orbit ISLs are just relayed without changing their direction. However, packets received through inter-orbit ISLs are forwarded according to the routing table in order to avoid needless delay, delay jitter and packet re-ordering. Note that if the routing table output is *u*, packets are forwarded to the successor (active counterpart), since the UDL links are inactive.

Table 5.2 illustrates the forwarding table for each state. Recall that interface names are defined in Figure 5.9 and I_{RT} stands for outgoing interface according to the routing table. Table 5.3 summarizes the MSVN-SHO operation. Durations of the states will be described later.

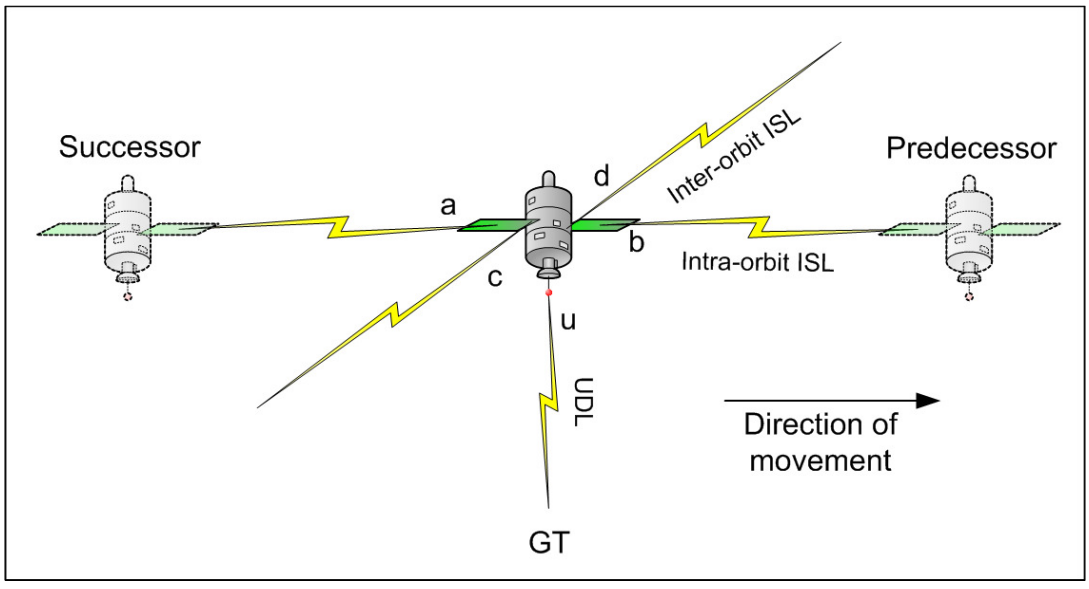


Figure 5.9. Link interfaces of a satellite

Table 5.2. MSVN-SHO forwarding table

<table>
<tr><td rowspan="2">Incoming Interface (I-Int)</td><td colspan="5">Outgoing Interface (O-Int)</td></tr>
<tr><td>Passive 1</td><td>Passive 2</td><td>Active 1</td><td>Active 2</td><td>Active 3</td></tr>
<tr><td>a</td><td>b</td><td>b</td><td>I_{RT}</td><td rowspan="3">I_{RT}</td><td>b</td></tr>
<tr><td>b</td><td>a</td><td>a</td><td>a</td><td rowspan="3">I_{RT}</td></tr>
<tr><td>c</td><td>b</td><td rowspan="2">I_{RT} (if I_{RT}=u then a)</td><td rowspan="3">I_{RT}</td></tr>
<tr><td>d</td><td>b</td></tr>
<tr><td>u</td><td>inactive</td><td>inactive</td><td></td><td></td></tr>
</table>

Table 5.3. MSVN-SHO operation

Step 1. Switch to new footprint area. *State* = passive 1, *duration* = t_{p1}.

Step 2. Receive routing table from predecessor.

Step 3. Update tables and activate UDL links. Send ACTIVE to predecessor and GTs. *State* = active 1, *duration* = t_{a1}.

Step 4. Receive INACTIVE from predecessor. *State* = active 2, *duration* = t_{a2}.

Step 5. Send routing table and state info to successor.

Step 6. Receive ACTIVE from successor. *State* = active 3, *duration* = t_{a3}.

Step 7. Deactivate UDL links. Send INACTIVE to successor. *State* = passive 2, *duration* = t_{p2}.

- For each state, packet forwarding is done according to Table 5.2.

Let us say two satellites (Sat1 and Sat2) fly over the same footprint area. Initially, Sat2 is in active mode and Sat1 is in passive mode. When GTs on the footprint area receive an ACTIVE message from Sat1, they start sending to and receiving from both of the satellites. Note that, when both of the satellites are in the active mode, packet duplication occurs. To eliminate duplication and losses, Sat2 and Sat1 should agree on a boundary (a packet number or a time unit), such that packets that arrive from GTs prior to the boundary are handled by Sat2 and ignored by Sat1. Similarly, packets that arrive from GTs after the boundary are handled by Sat1 and ignored by Sat2. In other words, if we satisfy that Sat1 switches to active mode exactly at the same time when Sat2 switches to passive mode, packet duplication or loss will be eliminated. However, achieving perfect synchronization is practically difficult and system may allow small amount of duplication.

Note that in case two satellites change their mode exactly at the same time, ACTIVE message is received from the successor after sending INACTIVE message to it. In that case, active 3 state is eliminated and ACTIVE message is received in passive 2 state. Therefore, in the beginning of the passive 2 state (before receiving ACTIVE message),

packets received from the successor should be forwarded according to the routing table. If the O-Int of these packets are UDL downlink, they can be sent back to the successor or (regarding that GTs are capable of communicating with more than one satellites at the same time) Sat2 could activate UDL downlinks only for those packets (that arrived from interface *a* before an ACTIVE message).

Let us define t_{a1}, t_{a2}, t_{a3}, t_{p1}, t_{p2} as duration of active (1,2,3) and passive (1,2) states respectively. In a perfect synchronization case as mentioned above, we can say that:

1. Passive duration of a satellite ($t_{p1} + t_{p2}$) is equal to $M_{\mathrm{FP}}^{\mathrm{L}} \cdot \Delta t_{\mathrm{ST}}$.
2. Active duration of a satellite ($t_{a1} + t_{a2}$) is equal to $M_{\mathrm{FP}}^{\mathrm{H}} \cdot \Delta t_{\mathrm{ST}}$.
3. t_{a1} is equal to intra-orbit ISL link delay (T_{ISL}).
4. t_{a3} is equal to zero.

If the new path is shorter than the old path, packet reordering could occur. New path is typically shorter when the Sat1 is closer to the other end node than Sat2. Expected delay difference is T_{ISL}. To avoid packet reordering, packets traversed over the new path can be delayed (T_{ISL} amount of time) at an appropriate intermediate or end node. On the other hand, if Sat2 is closer to the other end node, receiver experiences the same amount of delay between packets traversing the old path and the packets traversing the new path.

Note also that, in MSVN-SHO, perfect synchronization between satellites in different orbits is not a necessity for successful transmission of packets, but it is desirable for avoiding delay variation. Suppose that Sat1 and Sat3 are connected with each other with an inter-orbit ISL. If Sat3 switches to active mode while Sat1 stays in passive mode, then Sat1 directs the packets received from Sat3 to Sat2, and Sat2 forwards to appropriate node (may be to Sat1 again) looking at its routing table. This may lead to additional delay and delay variation for a flow. This situation can be eliminated if Sat3 and Sat1 (or any two satellites connected with inter-orbit ISLs) switch to the active (or passive) mode almost at the same time.

Recall that MSVN-SHO algorithm is designed for the case of $1< N_{S/F}^{avg} <2$. We considered this case because system cost becomes too high when $N_{S/F}^{avg}$ exceeds 2, and it is not reasonable to double the number of satellites in order to benefit from soft handover. Nevertheless, case of $N_{S/F}^{avg} >2$ could be considered for the sake of increasing system availability. In that case, there will be more than one active satellite with a single passive counterpart and proposed handover algorithms require some modification. According to the number of additional active satellites serving for a single footprint area, there will be additional active modes in the state diagram. When a satellite enters to a footprint area, it will be first in passive 1 mode, then switch to active modes corresponding to first active VN. Then it will switch to active modes corresponding to next active VN(s). Finally it switches to passive 2 mode and leave the footprint area. MSVN-SHO operation and the forwarding table should be extended accordingly. In this thesis, we skip the details of the modification.

5.4.2.1. MSVN-SSHO Algorithm. Basically MSVN-SSHO is very similar to MSVN-SHO. The satellite state diagram shown in Figure 5.8, Table 5.2 and Table 5.3 are also valid for MSVN-SSHO (except that active 3 state does not exist). It considers the case where GTs cannot receive/send different data from/to more than one satellite at the same time. Again, let us consider two satellites (Sat1 and Sat2) flying over a footprint area. Initially Sat1 is passive, and Sat2 is active and communicates with GTs over the footprint area. Since we consider that only one satellite serves for a footprint at any time, two satellites can use exactly the same channel assignment for communicating with GTs. Therefore, GTs are uninterrupted while switching from Sat2 to Sat1 (Such interruption avoidance is also considered in Teledesic [60]). Different from MSVN-SHO, GTs could not receive packets from or send packets to Sat2 after switching to Sat1. In such a case, some packets may needlessly shuttle between Sat1 and Sat2, i.e., when Sat1 is passive, it relays the packets (that are to be forwarded to downlink UDL) to Sat2, and when they reach at Sat2, Sat2 also becomes passive and resend them back to Sat1. (Note that in MSVN-SHO, such packets could be sent to GTs by Sat2, since MSVN-SHO is designed for the case that GTs are capable of communicating with more than two satellites at the same time.) In order to avoid this situation, Sat1 may buffer such packets starting from an

appropriate time before it switches to active state, and send to GTs after activating UDL link.

To avoid packet losses, time between Sat2's deactivating UDL link and Sat1's activating UDL link should be minimized. This necessitates perfect synchronization between Sat1 and Sat2.

5.4.3. Comparison of VN-HO, MSVN-SHO and MSVN-SSHO

Let us consider two polar satellite systems with 12 orbits and N_{FP} = 24. The first one is a VN-based satellite system with N_{SAT} = 24 and the other is an MSVN-based satellite system with N_{SAT} = 25. We assume that satellites are synchronized perfectly. For VN-based system, link delay values are similar to that considered in [61], such that T_{UDL} and T_{ISL} (intra-orbit) are 7 ms and inter-orbit ISL delay is 14 ms. For MSVN-based system these values are assumed to be same except that T_{ISL} is 6.72 ms since satellites in the same orbit are closer to each other. We assume that packet processing time is negligibly small compared to the link delay values.

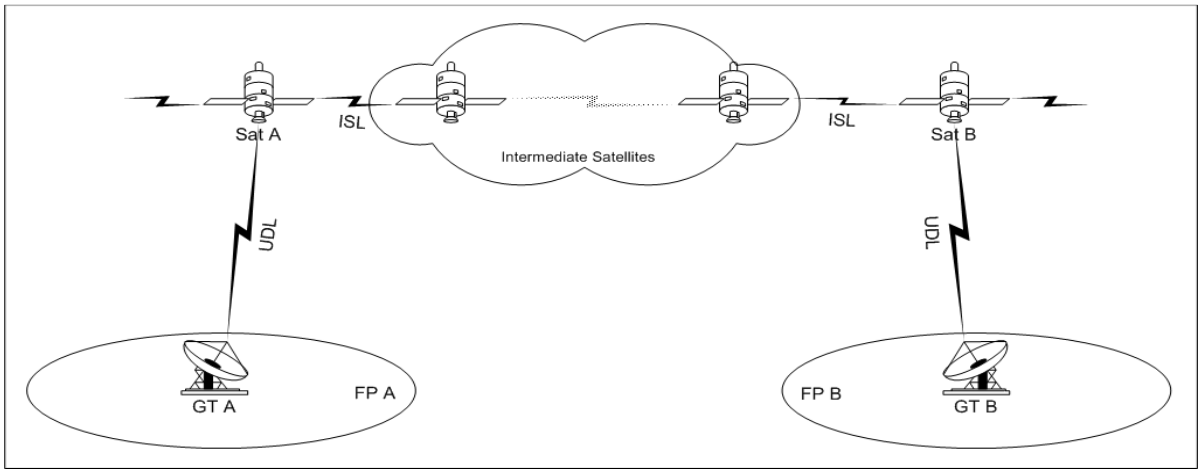


Figure 5.10. Communication scenario between two ground terminals.

For both systems, we consider part of the network as shown in Figure 5.10. We consider a flow between two ground terminals, GT-A and GT-B and investigate the handover performance for different instances of FP-A and FP-B. We assume that the flow

has constant bit rate of 10 Mbps with guaranteed QoS. Results show that handover performance is independent from number of hops between Sat-A and Sat-B for both VN-based system and MSVN-based system.

Table 5.4. Handover latency values for proposed handover mechanisms

	VN-HO	MSVN-SHO	MSVN-SSHO
T_{SW} = 10 ms	24 ms	No latency	No latency[2]
T_{SW} = 20 ms	34 ms		
T_{SW} = 50 ms	64 ms		
T_{SW} = 100 ms	114 ms		

Handover latency for VN-HO is dependent on T_{SW} as shown in Table 5.4. Obtained values are consistent with Equation 5.20. Latency is due to the fact that UDL transmissions stop in VN-based systems when satellites switch their footprint area. On the other hand, MSVN-based handover algorithms provide zero latency since UDL transmission does not stop. Recall that we assume perfect synchronization of satellites. If perfect synchronization could not be provided, latency for VN-HO and MSVN-SSHO would increase.

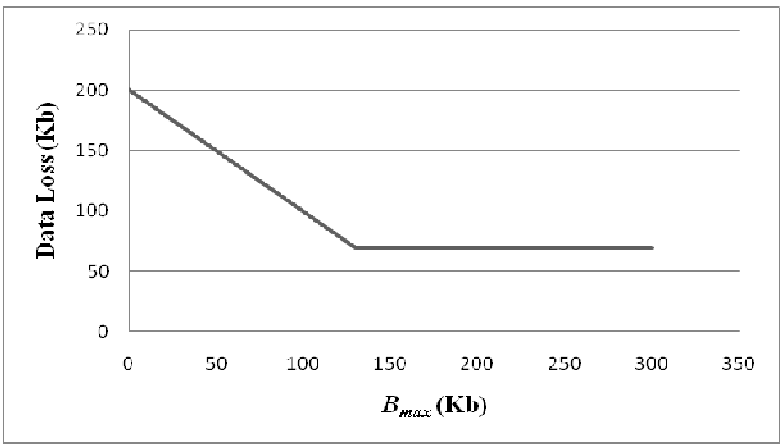


Figure 5.11. Data loss values for VN-HO algorithm

[2] Although MSVN-SSHO provides uninterrupt communication between GTs and satellites, small amount of latency may occur when switching from one satellite to another. In this work, we assume that this latency is negligible.

Figure 5.11 illustrates *data loss* values for VN-HO algorithm for different B_{max} values. B_{max} is defined as maximum amount of onboard buffering available for the corresponding flow. T_{SW} is set to 20 ms. Since T_{SW} is higher than T_{ISL}, buffering is needed to avoid data loss during switching of satellites from one footprint to another. If small amount of buffering is supported onboard (or buffering is not supported at all), data loss is inevitable. For high B_{max} values (higher than 130 Kb in this case), packets on the network can be buffered until the end of the switching time and only data loss occurs in uplink UDL (which is equal to 70 Kb). MSVN-SHO is a lossless handover algorithm. MSVN-SSHO also avoids data loss in the case of perfect synchronization. In MSVN-SSHO, perfect synchronization between successive satellites along an orbit is important, since both of them could not be active (for same footprint area) at the same time. If both of them are passive at the same time, then GTs will not be served and packet loss occurs. Therefore, for the best handover performance, satellites should change their mode in a synchronized way. Perfect synchronization is also needed for VN-HO to reduce the data loss (or latency if packets are buffered). In MSVN-SHO, perfect synchronization is not a prerequisite for avoiding packet loss. However, perfect synchronization between successive satellites along an orbit is beneficial for avoiding packet duplication, and synchronization between neighboring satellites in different orbits is desirable for avoiding needless delay variations (and hence packet reordering).

In VN-HO, *delay variation* in a flow occurs due to directing the packets from old representative of a VN to the new representative after handover occurs. As handover is realized, mapping between logical nodes and physical satellites changes and packets on the network should travel an ISL link in order to continue from their old logical node. In our scenario, 7 ms of end-to-end delay variation occurs at the time of handover. In MSVN-SHO, the mapping between logical and physical nodes changes in a smooth way. Before changing its logical position, a satellite stays in passive mode, and handles the traffic that is directed to itself in a proper way. Nevertheless, delay variation in flow occurs due to length difference between the old path and the new path. In our scenario, this corresponds to a delay variation of 0 ms and 6.72 ms for different instances of FP-A and FP-B. When FP-A and FP-B are in the same latitude, path length does not change, and hence end-to-end delay doesn't vary during handover. However, in other cases, path length varies at the amount of one intra-orbit ISL distance. When the new path is shorter, delay variation

causes packet reordering. As previously mentioned, packet reordering may be handled by delaying packets that travel along the new route in an appropriate node, if needed. Delay variation in MSVN-SSHO is also similar to MSVN-SHO, but additional variation occurs because just after a satellite switches to passive mode, some packets that arrived from successor should be sent back. Therefore, comparing to the MSVN-SHO case, two times more delay variation (13.44 ms) is obtained. This variation could be reduced by buffering those packets in the successor, if possible.

There is no significant difference between proposed handover algorithms in terms of *signaling complexity*. In VN-HO, a satellite receives and sends state information once in a system period. It sends total of four ACTIVATED messages to its neighbors (assuming that it has four neighbors), and send STOP and READY messages to GTs. In MSVN-based handover algorithms, again each satellite receives and sends RT and state information once in a system period. It sends one ACTIVE and one INACTIVE message to one of its neighbors, and to GTs. Therefore, signal generated from a single satellite seems to be slightly low in MSVN-based handover algorithms, but in overall, extra number of satellites should be taken into account.

From the above descriptions, it is evident that MSVN-based handover algorithms perform better than VN-HO. Especially MSVN-SHO performs very well in terms of handover latency, data loss, delay variation, and synchronization and buffering requirements. However, since VN-HO, MSVN-SHO and MSVN-SSHO are designed for different systems with different capabilities, system characteristics should be taken into account in order to make a better comparison.

MSVN-SHO is better than MSVN-SSHO in terms of delay variation and less need for synchronization and buffering. However, enabling GTs to communicate with more than one satellite at a time necessitates extra complexity in the system.

MSVN-based handover algorithms are faster and smoother than VN-HO, but MSVN-based systems use more satellites than VN-based systems. Actually, increase in the system cost is marginal since using N_{FP} +1 satellites per orbit is usually sufficient to apply

MSVN-based handover algorithms. Moreover, in the passive mode, satellites use small amount of power, and life-time of the satellites in an MSVN-based system increases comparing with a VN-based system. Therefore, increasing number of satellites to benefit from performance of MSVN-based handover algorithms seems to be a reasonable decision.

5.5. Optimal Beam Management

In previous section, we described high performance handover mechanisms as one of the possible advantages of MSVN-based systems. In this section, we are going to reveal another possible advantage of using such systems. As we described in Section 5.3.2, system availability of an MSVN-based satellite system may be increased by making some of the satellites to direct their beams to neighboring footprint areas. In other words, mapping between satellites and ground areas play an important role for increasing system availability and throughput. One should consider non-homogeneity of traffic distribution over the globe to support optimal resource utilization. In this subsection, we aim to provide optimal mapping between satellites and ground terminals and for this purpose we propose and evaluate a beam management technique where satellites over rural areas may be made to direct their beams to denser areas. Section 5.5.1 describes the considered problem and the solution approach, and Section 5.5.2 gives the numerical results.

5.5.1. Beam Management Problem

For the systems with non-integer $N_{\mathrm{S/F}}^{\mathrm{avg}}$ value, number of satellites serving an area switches between $\lfloor N_{\mathrm{S/F}}^{\mathrm{avg}} \rfloor$ and $\lceil N_{\mathrm{S/F}}^{\mathrm{avg}} \rceil$, and this leads to instability in the bandwidth supply offered by the system as described in Section 5.3.2. Nevertheless, system availability could be increased by properly adjusting beam directions of the satellites. In other words, by making satellites to direct their beams to neighboring footprint areas (with higher traffic density) we can increase the overall service offered by the satellite system.

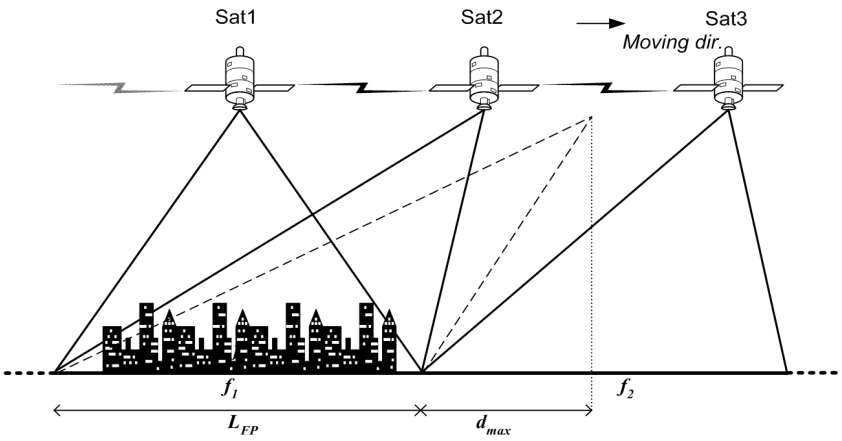


Figure 5.12. Directing satellite beams to dense areas

In the regular scheme, at a given time each satellite serves for the nearest footprint area, i.e. the one with higher elevation angle. When a satellite comes to a point that is closer to another footprint area, it switches serving that area. We call this point *regular switching point*. However, since traffic is non-homogenously distributed over globe, relaxing this rule could increase system throughput. Moreover, it renders possible increasing actual service as described in previous section. Figure 5.12 illustrates the task of directing satellite beams to neighboring areas. Since f_1 is denser than f_2, Sat2 continues serving f_1 and does not switch to serving f_2 as in the regular scheme. A satellite can serve its neighboring area until reaching a point, which is measured by angular distance $d_{\max}$ to the regular switching point. Let L_{FP} denotes angular length of a footprint. We define $r_{\max}$ as ratio of $d_{\max}$ to L_{FP} and in this work we assume that it does not exceed 1.

$$d_{\max} = r_{\max} \cdot L_{\text{FP}} \tag{5.21}$$

Let $B_{\text{SUP}}^{\text{sat}}$ is bandwidth supplied by a satellite, and B_{SUP} is the average bandwidth supplied to a footprint area in the regular scheme. It is clear that:

$$B_{\text{SUP}} = B_{\text{SUP}}^{\text{sat}} \cdot N_{S/F}^{\text{avg}} \tag{5.22}$$

For the sake of increasing system availability, satellites may assign bandwidth to neighboring areas. Maximum possible amount of directed bandwidth is indicated by $B_{\text{DIR}}^{\max}$.

$$B_{\mathrm{DIR}}^{\max} = B_{\mathrm{SUP}} \cdot r_{\max} \tag{5.23}$$

For the sake of simplicity, we assume that adjacent satellite signal interference is totally eliminated. $B_{\mathrm{DIR}}^{f_2,f_1}$ represents the amount of bandwidth directed from footprint f_2 to footprint f_1. It is between zero and $B_{\mathrm{DIR}}^{\max}$. Directing $B_{\mathrm{DIR}}^{f_2,f_1}$ amount of bandwidth from f_2 to f_1 means that each satellite continues serving f_1 until reaching a point which is d_{f_2,f_1} units (radians) far from the regular switching point between f_1 and f_2.

$$d_{f_2,f_1} = \frac{B_{\mathrm{DIR}}^{f_2,f_1}}{B_{\mathrm{SUP}}} \cdot L_{\mathrm{FP}} \tag{5.24}$$

Each footprint area has a bandwidth demand depending on the user density, etc. Bandwidth demand is denoted by B_{DEM}^{f}, where $f \in \{1,2,3,...,N_{\mathrm{FP}}\}$ represents footprint number. Because of the limited resources, system may not satisfy all of the demand. B_{STS}^{f} denotes the satisfied amount of demand for footprint f. We aim to maximize satisfaction of demands over the whole globe. Now, we will give a formulation for this problem.

5.5.1.1. Problem Formulation. We assume bandwidth direction is possible only between consecutive footprint areas that belong to same orbit. Hence, demand satisfaction problem can be handled for each orbit independently. Solving N_P maximization problems, one for each orbit plane, we come up with optimal beam management solution for the whole constellation. Therefore, for each satellite orbit, we give the following problem formulation:

$$\max \sum_{f} B_{\mathrm{STS}}^{f} \text{ where}$$

$$B_{\mathrm{STS}}^{f} \le B_{\mathrm{DEM}}^{f} \tag{5.25}$$

$$B_{\mathrm{STS}}^{f} \le B_{\mathrm{SUP}} + B_{\mathrm{DIR}}^{f\text{-}1,f} + B_{\mathrm{DIR}}^{f+1,f} - B_{\mathrm{DIR}}^{f,f-1} - B_{\mathrm{DIR}}^{f,f+1} \tag{5.26}$$

$$0 \le B_{\mathrm{DIR}}^{f_1,f_2} \le B_{\mathrm{DIR}}^{\max} \tag{5.27}$$

Note that, in constraint 5.26, we should replace $f-1$ with N_{FP} for $f=1$, and $f+1$ with 1 for $f = N_{FP}$. Actually the above problem can be considered as a maximum-flow problem as shown in Figure 5.13. Values over the edges correspond to upper bounds on arc capacities (lower bounds are zero). Aim is to send as much flow as possible form *S* to *T*. Problem can be solved by any polynomial time algorithm proposed for maximum flow problem [62].

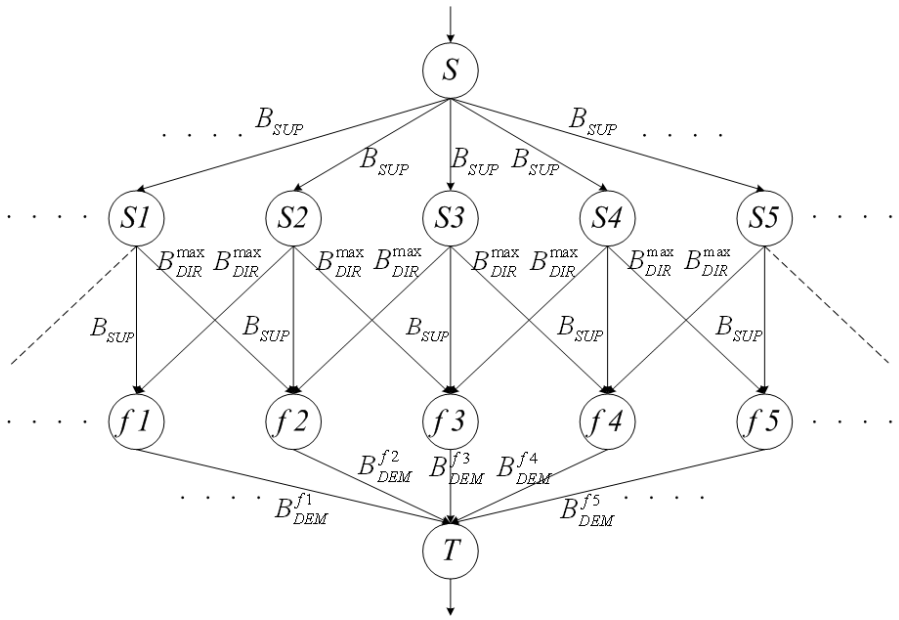


Figure 5.13. Illustration of the maximum flow problem

In the above formulation, while determining B_{STS}^{f}, we take account of the average bandwidth supplied to the footprint area. However, as we described in the previous section, it is more reasonable to take care about minimum bandwidth supplied (due to the instability reasons). Therefore we change constraint 5.26 as following:

$$B_{\mathrm{STS}}^{f} \le \left\lfloor \frac{B_{\mathrm{SUP}} + B_{\mathrm{DIR}}^{f\text{-}1,f} + B_{\mathrm{DIR}}^{f+1,f} - B_{\mathrm{DIR}}^{f,f-1} - B_{\mathrm{DIR}}^{f,f+1}}{B_{\mathrm{SUP}}^{sat}} \right\rfloor \cdot B_{\mathrm{SUP}}^{sat} \tag{5.28}$$

Moreover, we should include one more constraint to ensure that at least one satellite serves for each footprint area at any time, regardless of the traffic density:

$$B_{\text{SUP}} + B_{\text{DIR}}^{f\text{-}1,f} + B_{\text{DIR}}^{f+1,f} - B_{\text{DIR}}^{f,f-1} - B_{\text{DIR}}^{f,f+1} \geq B_{\text{SUP}}^{sat} \tag{5.29}$$

The new problem formulation is not linear any more. However, by modifying the network flow problem shown in Figure 5.13, we can come up with a feasible solution approach.

5.5.1.2. Solution Approach. Constraint 5.28 states that the utilizable amount of offered capacity is the integer multiples of $B_{\text{SUP}}^{\text{sat}}$. This suggests us to convert the maximum flow problem to feasible flow problem shown in Figure 5.14. ($K_i \cdot B_{\text{SUP}}^{\text{sat}}$, $K_i \cdot B_{\text{SUP}}^{\text{sat}}$) tuples stand for the lower bound and upper bound (which are same) for the corresponding arc. For each footprint area, we try to supply bandwidth that is integer multiple of $B_{\text{SUP}}^{\text{sat}}$. In other words, for footprint *f*, we try to send $K_f \cdot B_{\text{SUP}}^{\text{sat}}$. For each possible combinations of $\left(K_1, K_2, K_3, ..., K_{N_{\text{FP}}}\right)$ we find whether there is a feasible flow, and if there is, we call it *feasible K-combination*. Then, among all feasible *K*-combinations, we select the one which offers best satisfaction (satisfied amount of traffic demand).

Complexity of the solution algorithm is related to the number of possible *K*-combinations. In most scenarios, it is expected to be not so large due to several restrictions. Firstly, from constraint 5.29 any *K* value should be greater than one:

$$\forall f \;\; K_f \geq 1 \tag{5.30}$$

Moreover, *K* values are limited with the maximum amount of possible bandwidth supply.

$$\forall f \;\; K_f \cdot B_{\text{SUP}}^{\text{sat}} \leq B_{\text{SUP}}^{\text{avg}} + 2 \cdot B_{\text{DIR}}^{\max} \tag{5.31}$$

For any two consecutive footprints f_1 and f_2:

$$\left(K_{f_1} + K_{f_2}\right) \cdot B_{\text{SUP}}^{\text{sat}} \leq 2 \cdot B_{\text{SUP}}^{\text{avg}} + 2 \cdot B_{\text{DIR}}^{\max} \tag{5.32}$$

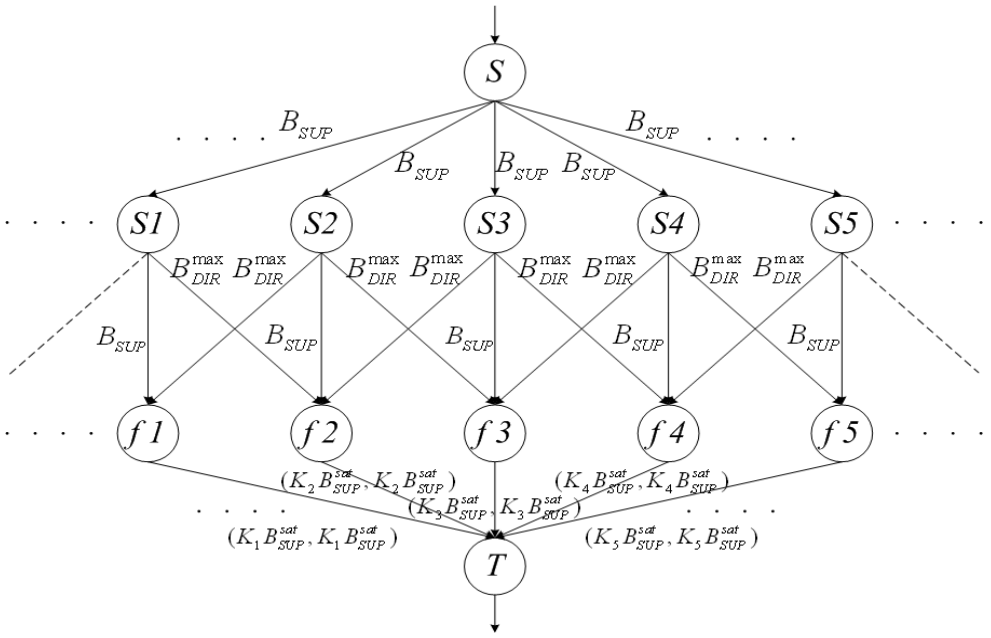


Figure 5.14. Illustration of the solution approach (feasible flow problem)

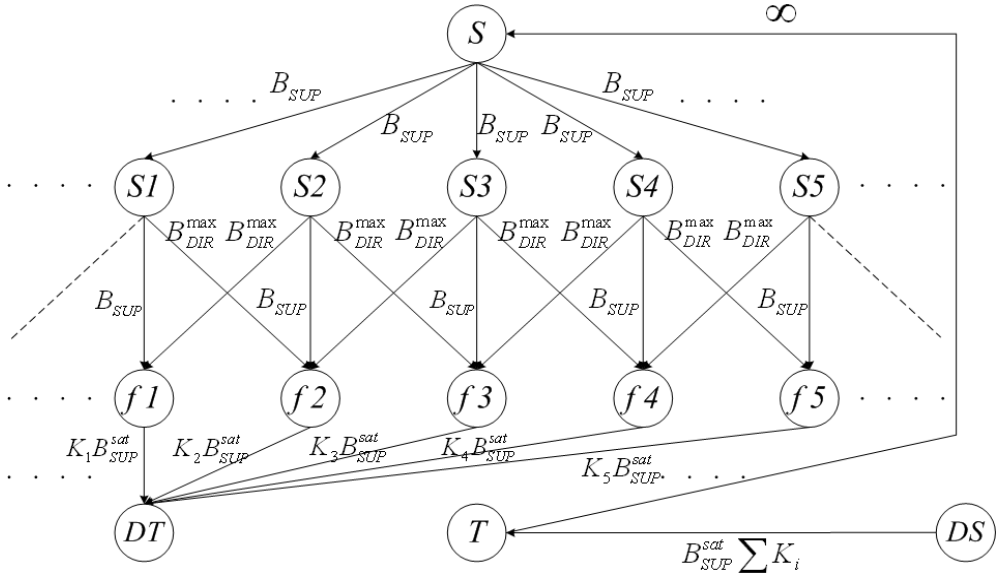


Figure 5.15. Transformed network flow problem

In general, for n consecutive footprints ($f_1, f_2, \ldots, f_n$):

$$\begin{aligned} \left(K_{f_1} + K_{f_2} + \ldots + K_{f_n}\right) \cdot B_{\mathrm{SUP}}^{\mathrm{sat}} &\le n \cdot B_{\mathrm{SUP}}^{\mathrm{avg}} + 2 \cdot B_{\mathrm{DIR}}^{\max} \quad \text{if } n < N_{\mathrm{FP}}, \\ \left(K_{f_1} + K_{f_2} + \ldots + K_{f_n}\right) \cdot B_{\mathrm{SUP}}^{\mathrm{sat}} &\le n \cdot B_{\mathrm{SUP}}^{\mathrm{avg}} \quad \text{if } n = N_{\mathrm{FP}} \end{aligned} \tag{5.33}$$

Another constraint comes from the fact that bandwidth utilization is limited by traffic demands:

$$\forall f \;\; K_f \cdot B_{\mathrm{SUP}}^{\mathrm{sat}} \le \left| B_{\mathrm{DEM}}^{f} \right| \tag{5.34}$$

In this work, we take constraints 5.30, 5.33 and 5.34 into account while determining possible set of *K*-combinations. It should be noted that by using additional appropriate constraints, set size could be further decreased. Moreover, by sorting *K*-combinations by their offered satisfaction, we can reduce the number of trials needed to find the feasible *K*-combination that offers best satisfaction.

Feasible flow problem shown in Figure 5.14 can be converted to maximum flow problem as follows [62]. We first transform the problem into a circulation problem by adding an arc (*T*,*S*) of infinite capacity. The original problem admits a feasible flow if and only if the circulation problem admits a feasible flow. Then we define supplies/demands $b(\cdot)$ at each node as follows.

$$b(i) = \sum_{j:(j,i)\in A} l_{ji} - \sum_{j:(i,j)\in A} l_{ij} \tag{5.35}$$

where *A* represents set of all arcs in the network and l_{ij} denotes lower bound of arc from node *i* to node *j*. Then we subtract l_{ij} from each l_{ij} and u_{ij} (upper bound of arc from node *i* to node *j*). Therefore, we remove arcs that have same lower bound and upper bound. Next we introduce two new nodes, a dummy source (*DS*) and a dummy sink (*DT*) node. For each node *i* with $b(i) > 0$, we add an arc (*DS*,*i*) with capacity b(i), and for each node *i* with $b(i) < 0$, we add an arc (*i*,*DT*) with capacity $-b(i)$. We refer to the new network as *transformed network* shown in Figure 5.15. Then we solve a max- flow problem from node *DS* to *DT* in the transformed network. If the maximum flow saturates all the source and sink arcs, problem has a feasible solution; otherwise, it is infeasible.

For solving maximum flow problem, we use a modified version of shortest augmenting path algorithm [62]. Shortest augmenting path algorithm always augments flow along a shortest path from the source to the sink in the residual network. In this thesis,

we will not describe details of this algorithm. The modification we employed is that, we first augment flow on the *direct paths*, i.e. on the paths *DS−T−S−Si−fi−DT*, where $1 \le i \le N_{\mathrm{FP}}$. This modification is employed to avoid unnecessary direction of beams to neighboring footprint areas.

For each possible *K*-combinations we solve the above feasible flow problem and find the *best K-combination* that offers best satisfaction. maximum flow problem from node *DS* to node *DT* in the transformed network.

5.5.2. Numerical Results

To test the effect of the proposed beam management technique, we consider a reference network, which is a polar satellite constellation with 12 orbits and 24 footprints per orbit. The Earth surface is divided into 24×12 zones as shown in Figure 4.2 and each zone stands for a footprint area. Footprint areas on the same longitude are served by satellites on the same orbit. We set traffic demands proportional to the user density levels given in Figure 4.2. We define aggregate demand T_{agg} which is equal to the total demand in the globe. Then traffic demand for footprint *f* of orbit *n* can be defined as follows:

$$B_{\mathrm{DEM}}^{f,n} = T_{\mathrm{agg}} \cdot \frac{u_{f,n}}{\sum_{i,j} u_{i,j}} \tag{5.36}$$

where $u_{i,j}$ stands for user density level for footprint *i* of orbit *j*. We assume that each satellite offers a capacity of 15 Gbps (which is slightly higher than the satellite capacity of 13.3 Gbps offered by Teledesic system [63]). We define satisfaction ratio (R_{STS}) as the performance metric. It is simply ratio of total satisfied demand to the total demand.

Figure 5.16 illustrates the numerical results for different $N_{\mathrm{S/F}}^{\mathrm{avg}}$ values and for different aggregate demands. $r_{\max}$ is set to 0.5. Results show that slight increase in $N_{\mathrm{S/F}}^{\mathrm{avg}}$ can achieve significant improvement in satisfaction ratio, especially in high traffic load. When we further increase $N_{\mathrm{S/F}}^{\mathrm{avg}}$, all the demands are satisfied with the expense of increased

system cost. Note that adding one satellite per orbit results in $1/N_{SAT}$ increase in number of satellites in the system. However actual increase in the cost is lower, since notable portion of satellite system production elements (e.g. design and software development) are independent of the number of satellites manufactured.

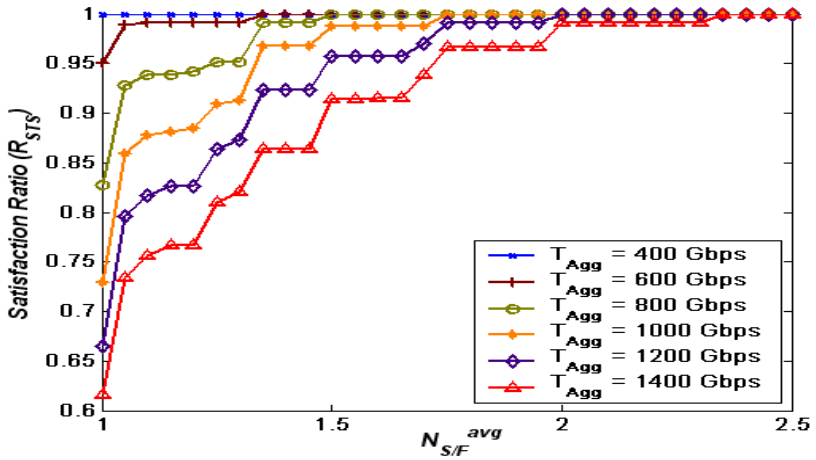


Figure 5.16. R_{STS} versus $N_{S/F}^{avg}$ for different traffic loads (r_{max}=0.5)

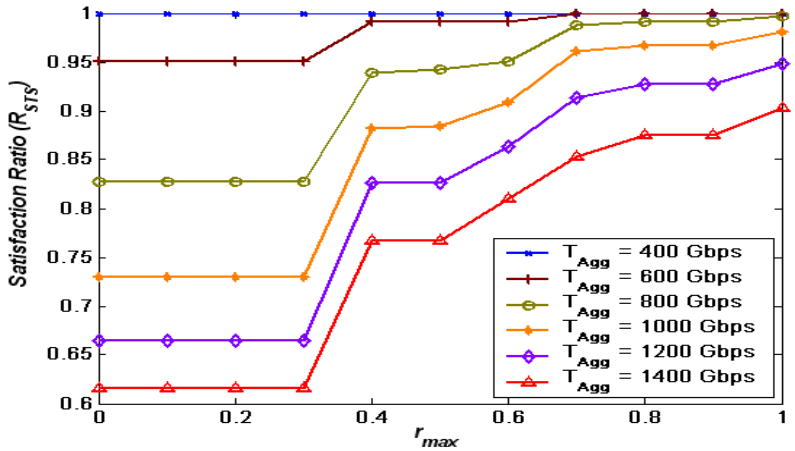


Figure 5.17. R_{STS} versus r_{max} for different traffic loads ($N_{S/F}^{avg}$ =1.2)

Next we keep $N_{S/F}^{avg}$ constant (at 1.2) and test the effect of r_{max}. Figure 5.17 shows that r_{max} has important effect on system throughput. For small values of r_{max}, system could not supply enough bandwidth to any of the footprint areas for increasing satisfaction ratio.

At the threshold value of 0.33, beam directing starts working for the scenario above. That is why there is a sudden increase in satisfaction at $r_{max} = 0.4$.

We evaluate contribution of beam management, assuming that demands are static. However, traffic demands are expected to vary at different time scales: daily, weekly and seasonal. To cope up with changes in traffic demands, a central node may gather information from the whole network periodically (or on demand), find the new optimal solution, and notify the relevant satellites to respect new beam direction pattern.

5.6. Summary

Handling satellite mobility is a major challenge for optimizing NGEO satellite network resources. Virtual Node (VN) based networking protocols were proposed for mobility handling, however these protocols require one-to-one correspondence between actual satellites and virtual nodes, resulting in reduced system availability. In this chapter, we investigate a general virtual topology for satellite systems with Earth-fixed footprints, where more than one satellite can serve for the same footprint area. We propose a multi-state virtual network (MSVN) topology, provide formal mathematical model for it and discuss its contribution to the overall system availability. Furthermore, we investigate potential handover mechanisms for VN-based and MSVN-based satellite systems, and propose efficient handover algorithms, namely VN-HO, MSVN-SHO and MSVN-SSHO. To our best knowledge, this is the first work to deal with handover algorithms in Earth-fixed satellite systems. Despite a marginal increase in the cost, MSVN-based systems offer handover algorithms that are faster and smoother than VN-HO. This constructs a significant benefit of MSVN-based satellite systems over conventional VN-based satellite systems.

After describing handover mechanisms, we focus on another approach to benefit from MSVN-based eath-fixed satellite systems. We propose an optimal beam management technique, which properly adjusts directions of satellite beams. Problem is formulated as a network flow problem, where aim is to maximize the satisfied demand. We test our contribution on a reference network and show that proposed beam management technique

could significantly increase system availability with slight increase in the system cost. Proposed optimization algorithm is performed in a central node, and it could be reperformed with proper intervals in order to adopt changes in traffic demands or satellite failures.

6. EFFICIENT INTEGRATION OF NGEO SATELLITE SYSTEMS WITH HIGH ALTITUDE PLATFORMS

6.1. Motivation and Related Work

High Altitude Platforms (HAPs) are aerial unmanned platforms operating in a quasi-stationary position at altitudes between 17 and 22 km [64]. Comparing with terrestrial infrastructure, they cover larger areas in line-of-sight propagation conditions. Moreover, they have advantages over satellites such as easy and incremental deployment, flexibility and reconfigurability, lower propagation delays and more favorable link budget of ground-HAP links. Therefore, HAPs are very suitable for providing last mile connectivity to the sensitive areas above where high bandwidth and accessibility are critical requirements. Moreover, HAPs are well suited for many mission-critical applications including real-time monitoring of seismic or coastal regions and terrestrial structures, real-time or non-real-time remote sensing and Earth observation for military or civic applications, pollution monitoring, traffic monitoring and control, and agriculture support, etc [64]. These HAPs are located over strategic areas, and receive large amount of data from terrestrial nodes and/or generate large amount of data using high-resolution optical and radar sensors. HAP stations transfer the received and generated data to Mission Control Centers (MCC) that are possibly distantly located and not in the coverage area of the HAP. Transmission of data from HAP to MCC can be done in three ways: 1) Via terrestrial links; 2) via multiple HAPs that are connected with inter-HAP links; 3) via satellites. The first case necessitates high-cost terrestrial infrastructure and it is not flexible. The second case necessitates existence of fully connected path of HAPs, but such a path may not exist in many cases where HAPs are not closely located. Transmission of data via satellite is the most feasible way.

With the advance of the free-space optical communications, HAPs and satellites can communicate with each other with large data rates. Recently, internetworking between satellites and HAPs are studied by various researchers. Several architectures consisting of terrestrial, HAP and GEO satellite layers are presented in several papers [64, 65, 66, 67]. In

such architectures, HAPs act as relay stations to aggregate and forward traffic received from the ground user terminals to the GEO satellite. GEO satellites are advantageous for their large coverage areas and static position with respect to the Earth, but they suffer from long propagation delays and high free-space attenuation. Optical links from HAPs to GEOs have less capacity (comparing to lower orbiting satellites) due to long distances [68], and long propagation delay is not suitable for real-time and interactive applications. LEO and MEO satellites offer much less propagation delays compared to GEO satellites and are attractive option for routing dense and real-time traffic. Therefore they are more suitable for emerging high data-rate and real-time mission critical applications. The most important challenge with the low orbiting satellites is their mobility with respect to the Earth, which complicates internetworking between these satellites and static nodes. Therefore, it is necessary to come up with solution approaches for this internetworking problem in physical layer and upper layers. In the literature, there are several studies that deal with internetworking between HAPs and LEO or MEO satellites. Most of these studies deal with physical layer issues such as fast pointing acquisition and tracking, handling the effect of the Doppler shift, optical transmitter design, etc. [69, 70, 71]. To our knowledge, there is almost no work that deals with resource management and routing issues related to upper layers. [30] focuses on routing VoIP traffic in a multilayered architecture with GEO satellites acting as the backbone routers, LEO satellites as the second layer and HAPs deployed in specific local regions. They simplify the mobility of the LEO layer by assuming that the physical LEO satellite network can be reduced to a fixed logical topology. As the authors indicate, this assumption is valid for the satellite systems with Earth-fixed footprints. However this is impractical in most satellite systems, and mobility of the satellites should be handled in more realistic ways.

In this chapter, we consider internetworking between HAPs and NGEO satellites without discarding the mobility of satellites. We consider an integrated architecture for mission-critical networking as described in the next section. HAPs and satellites communicate via high capacity free-space optical links. Each satellite can serve multiple HAPs and a HAP can have a line of sight with multiple satellites. Deciding on which satellite(s) to establish an optical link is an important issue. For instance, if the capabilities of the system allow single satellite to communicate with any HAP, the one with the higher elevation angle is the most appealing. This is because the higher the elevation angle, the

shorter is the link distance through the atmosphere. This leads to a reduced free space loss, a smaller atmospheric loss due to absorption and scattering, less background noise due to blue sky, and less fading. However, if we always make the HAP to communicate with the satellite with the highest elevation angle, link duration times will get shorter and switching (which is an expensive task) between HAPs and satellites occurs more frequently. Furthermore, number of optical receivers/transmitters in satellites is limited and this causes another constraint on the problem of linking satellites with HAPs. In this chapter, we consider all of the issues raised above and study on optimal matching of HAPs and satellites considering the movement of satellites and self-rotation of the Earth.

The rest of this chapter is organized as follows. In Section 6.2, we describe the system architecture and give mathematical formulas for visibility conditions, elevation angles and time-dependent locations of mobile satellites. Then we formulate the optimal link establishment problem after defining system constraints and objectives in Section 6.3. We also give a polynomial-time optimization algorithm for solving the formulated problem. Section 6.4 exhibits numerical results of the optimization algorithm performed for given integrated network scenarios, and Section 6.5 concludes this chapter.

6.2. System Overview

6.2.1. System Architecture

We consider a system with three layers: Ground layer, HAP layer and satellite layer. HAPs are located over sensitive and strategic areas, and receive and generate large amount of mission-critical data. They transmit data to low or medium orbiting satellites which act as relay nodes to transfer data to Mission Control Centers (MCC). If the corresponding MCCs are far away from HAPs, data could be sent over multiple satellite relays using inter-satellite links. The system architecture is shown in Figure 6.1. Due to the global coverage of satellites, this architecture enables world-wide mission critical networking.

HAPs and satellites communicate via free-space optical links. Free-space optical links have highly directive beams with very small divergence angles due to the short (near-

infrared) wavelength. As a result optical systems are extremely power efficient than microwave (MW) communications systems over long distances and allow data rates of several Gbps. Even faster data links would be feasible in the near future deploying methods like Wavelength Division Multiplexing (WDM). Optical links offer significantly higher bandwidth, however they are blocked by clouds. This does not impose a problem between HAPs and satellites because HAPs are located above the cloud layer. However, optical downlinks from satellites to MCCs have limited availability depending on the cloud situations. Since almost hundred percent availability is required for most mission-critical applications, using optical link from satellite to MCC is not feasible. On the other hand, MW communication systems are inhibited by spectrum restrictions, manageable antenna sizes, and available transmit power. Therefore for increasing downlink capacity, [68] proposed to use HAP relays between satellite and ground stations. Satellite transmits data to HAP via high-capacity optical link which is not hindered by clouds. The final "last mile" to the MCC could then be bridged by a standard point-to-point MW link as used today in terrestrial applications, but with a large bandwidth compared to a satellite link due to the short distance. In cloud-free conditions, a parallel optical link can be utilized to increase the bandwidth between HAP and MCC as shown in Figure 6.1. Moreover, a network of HAP interconnected with optical links can eliminate the cloud blockage problem by providing optical HAP-ground links at different geographic locations.

In this thesis, we leave details of downlinks from satellites to MCCs and will focus on establishment of optical links between HAPs serving for sensitive areas and mobile satellites. Each satellite can communicate with multiple HAPs, and each HAP can have a line-of-sight with multiple satellites. However, maintaining an optical link between a satellite and a HAP is a power consuming task. Since resources are restricted in satellites, there is an upper limit for number of HAPs to serve. Moreover, visibility conditions and elevation angles between satellites and HAPs continuously change due to mobility of satellites. Therefore, optimal matching of HAPs and satellites is a crucial issue, which will yield maximization of system availability and performance.

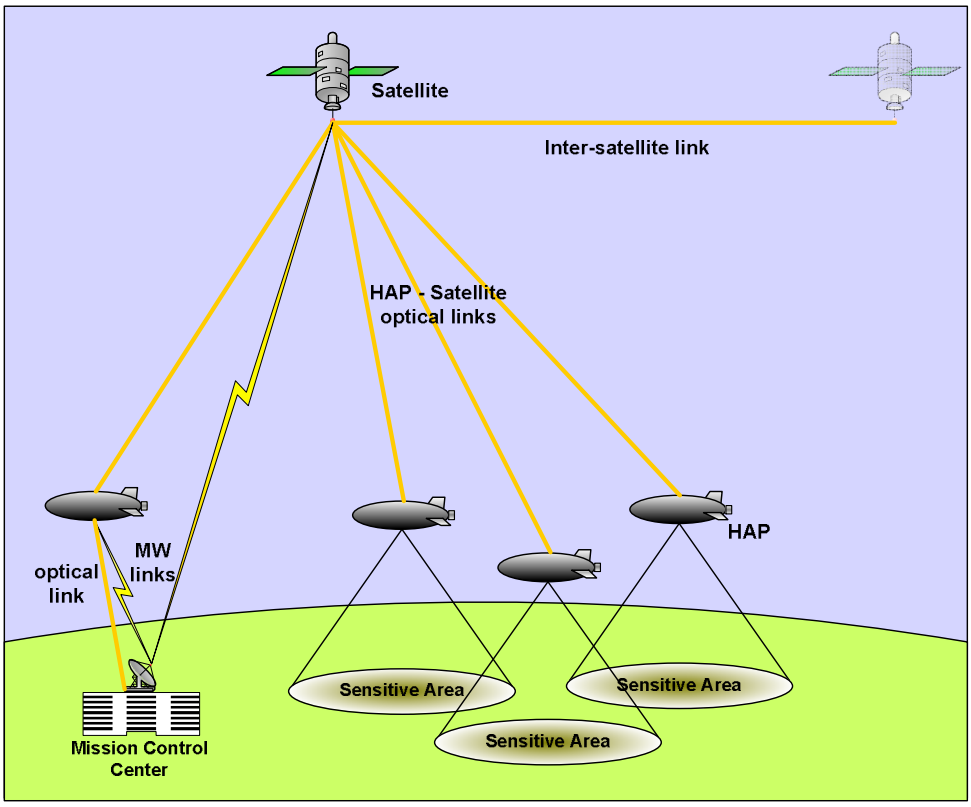


Figure 6.1. System Architecture

6.2.2. System Geometry

Figure 6.2 illustrates two-dimensional view of the considered system. There are three layers: Terrestrial layer, HAP layer and satellite layer. A HAP H is said to be visible to a satellite S if the elevation angle between them exceeds minimum elevation angle ($\varepsilon_{\min}$). This implies that it is possible to establish an optical link between a HAP and a satellite only if β does not exceed δ.

Applying the law of sines to sides OA and OS

$$\frac{\sin(90-\varepsilon_{\min}-\delta)}{R_{\mathrm{E}}+h_H}=\frac{\sin(90+\varepsilon_{\min})}{R_{\mathrm{E}}+h_S} \tag{6.1}$$

where R_{E} is the radius of the Earth (6375 km), h_H is the height of the HAP, and h_S is the height of the satellite. Extracting δ from Equation 6.1, we get

$$\delta = 90 - \varepsilon_{\min} - \arcsin\left(\frac{R_E + h_H}{R_E + h_S} \cdot \cos(\varepsilon_{\min})\right) \tag{6.2}$$

In Figure 6.2, *S'* is the projection point of the satellite *S* on the HAP layer. OS'H is an isosceles triangle, and again by using law of sines, β is found to be

$$\beta = 2 \cdot \arcsin\left(\frac{|S'H|}{2 \cdot (R_E + h_H)}\right) \tag{6.3}$$

It is possible to establish an optical link between a satellite and a HAP while $\beta \leq \delta$. β angle for a satellite-HAP pair continuously changes due to the movement of the satellites. For each time unit, we prepare a visibility matrix that represents which HAPs are visible to which satellites, based on the exact positions of the satellite and HAPs.

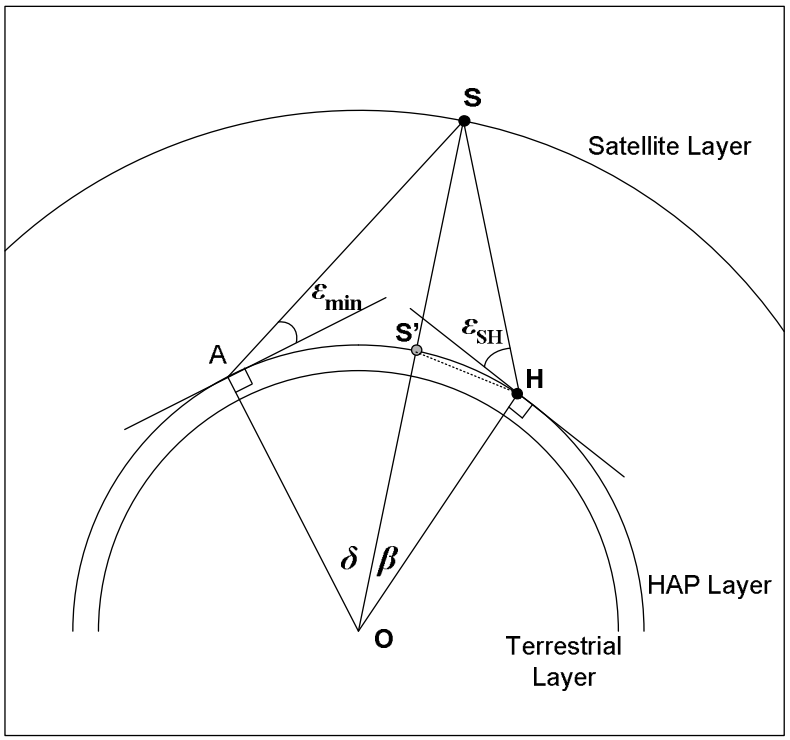


Figure 6.2. Two dimensional view of the system geometry

At a given time, the elevation angle between a satellite *S* and a HAP *H* (ε_{SH}) is

$$\varepsilon_{SH} = \arctan\left(\frac{1}{\sin(\beta)} \cdot \left(\cos(\beta) - \frac{R_{\mathrm{E}} + h_H}{R_{\mathrm{E}} + h_S} \right) \right) \tag{6.4}$$

which is derived by applying law of sines to OH and OS sides of the OSH triangle.

Obviously, β and ε_{SH} values are dependent on the locations of satellites and HAPs. Locations of HAPs are considered to be static, but this is not the case for satellites. Now, we are going to model the mobility of the satellite network.

Let us consider a regular satellite network with total number of $N = N_{\mathrm{P}} \times N_{\mathrm{S}}$ satellites where N_{P} is the number of orbit planes and N_{S} is the number of satellites per plane. A satellite is denoted by $S_{p,s}$, where p=1…N_{P}, s=1…N_{S}. Location of $S_{p,s}$ at time t is represented by ($\lambda_{p,s}(t)$, $\varphi_{p,s}(t)$) where $\lambda_{p,s}(t)$ is it's latitude value and $\varphi_{p,s}(t)$ is it's longitude value at time t. Assuming that location of $S_{1,1}$ at initial time t_0=0 is (0˚, 0˚), location of a satellite in orbit plane 1 at any given time t can be found as following :

$$\begin{aligned} \lambda_{1,j}(t) &= \arcsin(\sin\mu \cdot \sin\alpha_{\mathrm{I}}) \\ \varphi_{1,j}(t) &= \begin{cases} 360k_1 + \arccos(\cos\mu / \cos(\lambda_{1,j}(t))) - \omega_E t, \; if \; \lambda_{1,j}(t) \geq 0 \\ 360k_2 - \arccos(\cos\mu / \cos(\lambda_{1,j}(t))) - \omega_E t, \; if \; \lambda_{1,j}(t) < 0 \end{cases} \end{aligned} \tag{6.5}$$

where $\mu = \omega_S t + (j-1) \cdot \Delta\lambda$ with ω_{S} is the angular speed of a satellite, and $\Delta\lambda = 360 / N_{\mathrm{S}}$; α_{I} is the inclination angle of the satellite orbit; ω_{E} is the angular speed of the Earth; k_1 and k_2 are appropriate integers that satisfy $-180° \leq \varphi_{1,j} \leq 180°$.

Longitude and latitudes of the satellites on other orbit planes can be determined with respect to the satellites in the first orbit plane.

$$\begin{aligned} \lambda_{i,j}(t) &= \lambda_{1,j}(t + (i-1) \cdot \Delta t) \\ \varphi_{i,j}(t) &= \varphi_{1,j}(t + (i-1) \cdot \Delta t) + (i-1) \cdot (\Delta\varphi + \omega_{\mathrm{E}} \Delta t) + 360k \end{aligned} \tag{6.6}$$

where $\Delta\varphi = 180/N_P$ for π-constellations, and $\Delta\varphi = 360/N_P$ for 2π-constellations; $\Delta t = T_{sat} \cdot (\Delta\phi/360)$ where T_{sat} is the rotation period of satellites, and $\Delta\phi$ is the phase difference (in terms of degrees) between adjacent orbits; and *k* is an appropriate integer that satisfies $-180° \le \varphi_{i,j} \le 180°$.

6.3. Optimal Link Establishment

In this work, we focus on establishment of optical links between HAPs and satellites. Our main constraints are the following:

1. A satellite and a HAP should have line of sight in order to communicate with each other. This implies that elevation angle between a HAP and a satellite should exceed ε_{min}.
2. Number of optical transmitters/receivers in satellites is limited mainly due to the power limitations. In other words, a satellite *i* can serve maximum of H^i_{max} HAPs.
3. In this work, we consider one-to-many relation between HAPs and satellites. In other words, for each HAP, we consider establishing link with a single satellite.

Our aim is to match the satellites and HAPs in such a way that:

1. Above constraints should be satisfied.
2. As much HAP as possible should be served. In other words, utilization of the HAPs[3] (U_H) should be maximized. It should be 1 if it is possible. Utilization at a given time *t* is represented by U^t_H.

[3] Utilization of the HAPs is calculated as ratio of the number of served HAPs to the total number of HAPs.

3. Average of the elevation angles (A_{avg}) between satellites and HAPs will be maximized. A_{avg} at a given time t is represented by A_{avg}^{t}.

Note that, our most important goal is to maximize U_H (to serve as much HAP as possible). Among the satellite-HAP matchings that maximizes U_H, the best matching is the one that maximizes A_{avg}.

6.3.1. Problem Formulation

Suppose that our considered system includes N satellites and M HAPs. Firstly, we will define three matrices that represent relationship between these satellites and HAPs in a given time unit t.

1. $V_{N\times M}^{t}$ is the *visibility matrix*. If HAP j is visible to satellite i, then $V^t[i][j]=1$, otherwise $V^t[i][j]=0$. Visibility matrix is filled according to the relationship between β and δ as described in the previous section.
2. $EA_{N\times M}^{t}$ is the *matrix of elevation angles*. If $V^t[i][j]=1$, then $EA^t[i][j]$ stores the elevation angle between satellite i and HAP j. It is calculated by the Equation 6.4. If $V^t[i][j]=0$, then $EA^t[i][j]$ is set to a constant number Σ. We set Σ to a large negative number (such as -10,000).
3. $E_{N\times M}^{t}$ is the *existency matrix*. If optical link between satellite i and HAP j exists, then $E^t[i][j]=1$, otherwise $E^t[i][j]=0$.

Now, *optimal link establishment* problem can be formulated as following:

$$\max \sum_{i,j} E^t[i][j] \cdot EA^t[i][j] \tag{6.7}$$

subject to

$$\sum_{j} E^t[i][j] \le H^i_{\max}, \quad \forall i \tag{6.8}$$

$$\sum_{i} E^t[i][j] = 1, \quad \forall j \tag{6.9}$$

The objective is to match HAPs and satellites at a given time t such that sum of elevation angles for the connected HAP-satellite pairs is maximized. First constraint ensures that a satellite i can serve maximum of $H^i_{\max}$ HAPs. Second constraint implies that each HAP is matched to a single satellite.

In the above problem formulation, $E^t_{N\times M}$ is the only variable. Hence, the formulated problem is a binary integer linear programming (ILP) problem. Performing the optimization, we obtain the optimal E^t values. According to Equation 6.9, every HAP is matched with a satellite. However, some of the HAPs may not be able to be served due to resource restrictions. In that case, the above ILP forces them to match with a satellite that is not visible to it (which we call *void matching*). Therefore, we should update the E^t values. A HAP j is said to be not served by any satellites, if for any i, $E^t[i][j]=1$, but $V^t[i][j]=0$. Therefore, we perform the following operation to the obtained E^t values:

$$E^t \leftarrow E^t \cdot V^t \tag{6.10}$$

This operation nullifies void matchings obtained by the above ILP. After performing (10), total number of served HAPs (M_S) can be found as:

$$M_S = \sum_{i,j} E^t[i][j] \tag{6.11}$$

Hence, U^t_H can be calculated as:

$$U^t_H = \frac{M_S}{M} = \frac{\sum_{i,j} E^t[i][j]}{M} \tag{6.12}$$

As we mentioned before, our primary goal is to maximize the utilization. Note that the number of unserved HAPs (M_U) is equal to the number of void matchings. Therefore, maximizing U_H^t implies minimizing the number of void matchings. In our problem formulation, this goal is achieved by setting Σ to a large negative number. Recall that, for void matching between satellite i and HAP j, $V^t[i][j]=0$ and therefore $EA^t[i][j]$ is equal to Σ. Hence, the objective function 6.7 is equivalent to:

$$\max \quad EA_{SUM}^t + M_U \cdot \Sigma \tag{6.13}$$

where EA_{SUM}^t represents the sum of elevation angles for valid satellite-HAP matchings. Since Σ is a large negative number, it is apparent that our ILP formulation involves minimization of M_U value (in other words maximization of U_H value).

$M_U \cdot \Sigma$ is equal for all possible matchings with the same U_H^t value. Therefore, the objective function 6.13 also implies our second goal, that is the maximization of EA_{SUM}^t (hence A_{avg}^t value) among the satellite-HAP matchings that maximize U_H^t. Note that A_{avg}^t value can be found as

$$A_{avg}^t = \frac{EA_{SUM}^t}{M_S} \tag{6.14}$$

At this point, we proved that the above problem formulation is fully appropriate for the optimal establishment of the optical links between HAPs and satellites. Now let us focus on the solution approach for the optimization problem. As we mentioned above, the problem formulation is a binary ILP problem. Binary ILP problems can be solved by some sort of optimization algorithms such as branch and bound algorithms. However, these algorithms have exponential time complexity, and applying them may result in excessively long time, especially for systems with large number of HAPs and satellites. Therefore, we come up with a solution approach with polynomial time complexity. In the next sub-section, we explain the details of the proposed solution approach.

6.3.2. Polynomial-time Solution Approach

We represent the integrated HAP-satellite system as a bipartite graph $G = (V_S, V_H, E)$ whose vertices can be divided into two disjoint sets V_S and V_H. V_S includes $H^i_{\max}$ nodes for every satellite i ($1 \le i \le N$), and V_H includes single node per HAP. Nodes in the set V_S are represented by $S_{i,h}$ (where $1 \le i \le N$, $1 \le h \le H^i_{\max}$ and $S_{i,h}$ is the h^{th} node for satellite i), and the nodes in the set V_H are represented by H_j (where $1 \le j \le M$, and H_j corresponds to HAP j). At a given time t, if $V^t[i][j]$=1, i.e. if satellite i and HAP j are visible to each other, then there exists an edge between H_j and every node corresponding to satellite i ($S_{i,1}$, $S_{i,2}$, ... $S_{i,H\max}$). The weight of each link is equal to $EA^t[i][j]$ as shown in Figure 6.3. If $V^t[i][j]$=0, then H_j does not linked to any of the nodes corresponding to satellite i. Figure 6.4 illustrates the corresponding bipartite graph G for a small sample system with four HAPs and two satellites with $H_{\max}$=2. In this example, HAP 2 is visible to both satellites, and the other HAPs are visible to a single satellite. Weights of the edges correspond to the angles of elevations between HAPs and satellites.

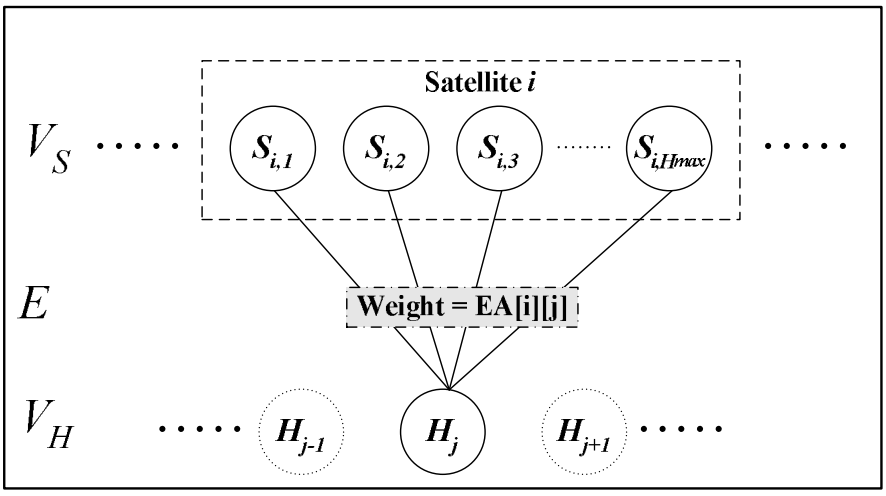


Figure 6.3. Bipartite graph representation of the system

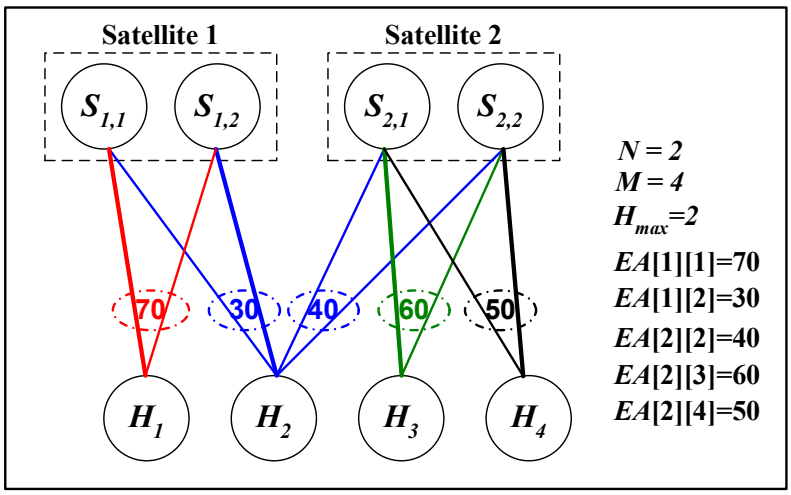


Figure 6.4. Bipartite graph representation of a small sample system with two satellites and four HAPs

Now, the formulated optimal link establishment problem is identical to the *maximum-weighted maximum-cardinality (MWMC) matching* in the constructed bipartite graph. In the mathematical discipline of graph theory, *matching* in a graph is defined as a subset of edges, such that no two edges share a common node. *Maximum-cardinality matching* is a matching with maximum number of edges. *MWMC matching* is a maximum cardinality matching such that the sum of the weights of the edges in the matching is maximum. In Figure 6.4, thick edges compose MWMC matching in the corresponding bipartite graph.

MWMC matching in the constructed bipartite graph satisfies all of the constraints and aims of the optimal link establishment problem. A satellite i can be assigned to maximum number of $H^i_{\max}$ HAPs, and a HAP can be matched with a single satellite, since there are $H^i_{\max}$ nodes for each satellite, a single node for each HAP, and no two edge can share a common node. Maximum possible utilization of HAPs (U_{H}) is achieved in the resulting matching, since it is a "maximum-cardinality matching", i.e. matching with maximum number of edges. A_{avg} value is also maximized, because sum of the weights of the edges in the matching is maximum.

In this section, we give a solution approach for MWMC matching problem based on the Hungarian Algorithm that is first developed by Kuhn [72]. Hungarian algorithm is an

effective polynomial-time combinatorial optimization algorithm for solving *linear assignment problem* which is defined as finding maximum-weighted matching in a balanced and complete bipartite graph. The algorithm finds a perfect matching, where each node in one partition is matched with exactly one node in the other partition, such that the sum of the weights of the edges in the matching is maximum. With the improving modifications applied to its initial version, it is proved that Hungarian algorithm has O($|V|^3$) time complexity in worst-case, where $|V|$ is the number of vertices [73].

Before applying the Hungarian algorithm, we should modify our bipartite graph *G* to a balanced and complete bipartite graph. Note that *G* is not complete, because there is no edge between H_j and nodes corresponding to satellite *i*, if HAP *j* and satellite *i* are not visible to each other. Therefore, we insert those missing edges and set their weights to Σ, which is a large negative number as defined in the previous section. The resulting graph is complete, but it is still not balanced, i.e. size of V_S is not equal to size of V_H. Therefore we introduce $\|V_S| - |V_H\|$ dummy nodes to make the graph balanced. We add dummy edges from these nodes to each node in the other partition and set their weights again to Σ. In the resulting bipartite graph, we run the Hungarian algorithm and find the maximum weighted perfect matching. Excluding all dummy edges from the resulting matching, we get a MWMC matching. In the resulting MWMC matching, if a node corresponding to satellite *i* is matched to a node corresponding to HAP *j*, then we set $E^t[i][j]$ to one. The resulting E^t matrix is identical to the output of the ILP formulated in the previous subsection. (Note that 6.10 is already performed and void matchings are already nullified by excluding dummy edges). Thus, we obtain optimal assignment of optical links that maximizes U_{H}, as well as A_{avg}.

6.3.3. Overall Optimization

Up to now, we consider the optimal matching issue for a given time *t*. However, since the satellites are mobile, connectivity conditions and elevation angle values between HAPs and satellites change with time. Therefore, optimal matching will change as time changes, and the optimization algorithm should be applied repeatedly in both periodic

manner (every Δt time units) and event-driven manner (when a link between a satellite and a HAP becomes obsolete, or a new HAP is joined to the system).

Note that satellite network topology is periodic, that is, it repeats itself within a known period. If the rotation period of satellites is T_{sat}, then the whole system period is found as

$$T_S = lcm(T_{\text{sat}}, T_{\text{E}}) \tag{6.15}$$

where T_{E} is the self-rotation period of the Earth, which is 24 hours. As long as HAP layer does not change, the whole system will repeat itself at each T_S time duration. Therefore, applying the optimization algorithm for a system period (T_S) will be enough as long as HAP layer remains static. If any HAP is removed from the system, joined to the system, or relocated, optimization process should be restarted. The overall optimization process is illustrated in Figure 6.5.

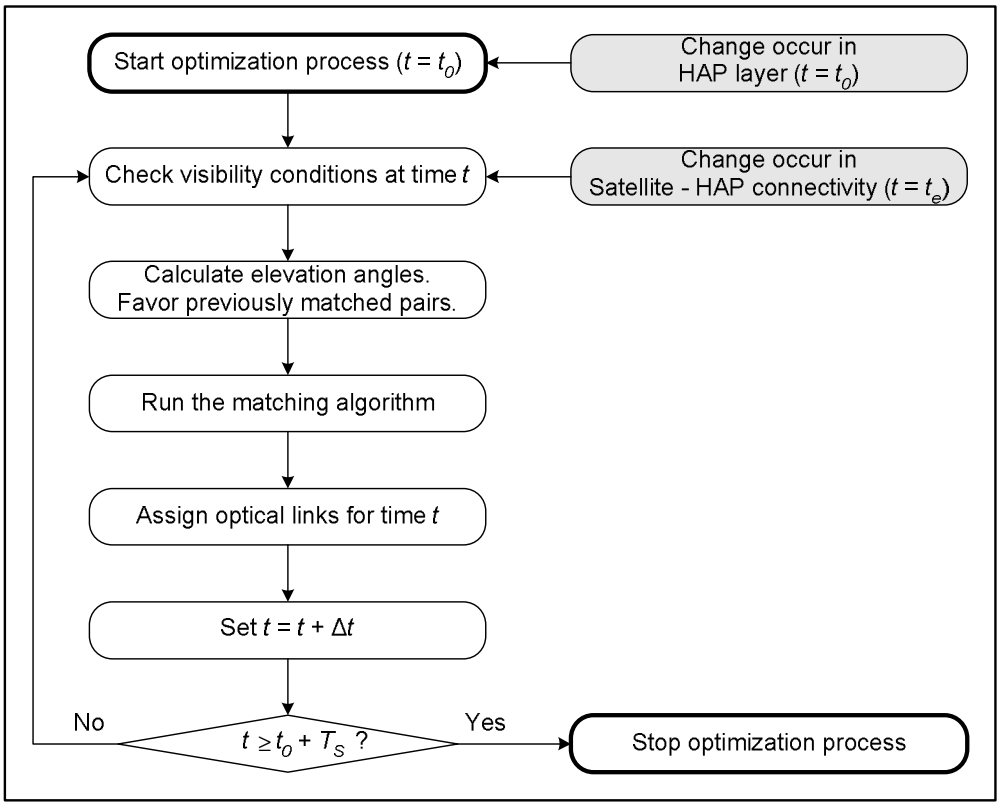


Figure 6.5. State diagram of the overall optimization process

While applying the optimization algorithm, we should also consider link duration times. Switching of some HAP-satellite links may result in small gain in A_{avg} value for a short time, but this gain may not compensate cost of the switching operation. Therefore, it is better to avoid establishment of links with a small duration, while aiming to maximize elevation angles. For this purpose, we propose to favor existing links in the optimization algorithm.

Consider that t_1 and t_2 are consecutive time instances. If a link exists between satellite i and HAP j at time t_1 then we increment $EA^{t_2}[i][j]$ value by a particular amount γ while applying the optimization algorithm for t_2:

$$if \quad E^{t_1}[i][j] = 1, \ then \quad EA^{t_2}[i][j] \leftarrow EA^{t_2}[i][j] + \gamma \tag{6.16}$$

γ value should be assigned appropriately. If γ is set to zero, then the optimization algorithm ignores link duration times, and aims to maximize A_{avg}. On the other hand, if γ is set to a large value (such as 90), then elevation angle is considered only when a link between a HAP and satellite is deactivated, and the algorithm has to decide to a new link. Once a link between a satellite and a HAP is activated, it will probably stay active until two nodes become invisible to each other, without caring the elevation angle degradation. In this work, we also investigate the effect of γ parameter, for achieving the optimal system performance.

6.4. Numerical Results

We simulate the optimal link establishment process in two different systems. First system (System-1) consists of a MEO satellite network that is identical to ICO constellation, and a hundred HAPs that serve for different sensitive areas distributed over globe as shown in Figure 6.6. In the second system (System-2), a LEO satellite network that is identical to Globalstar constellation serves for the same HAPs as shown in Figure 6.7. In both figures, dots represent static positions of HAPs and stars represent initial positions of satellites. HAPs are assumed to operate at 20 km altitude. Characteristics of ICO and Globalstar systems are given in Table 6.1.

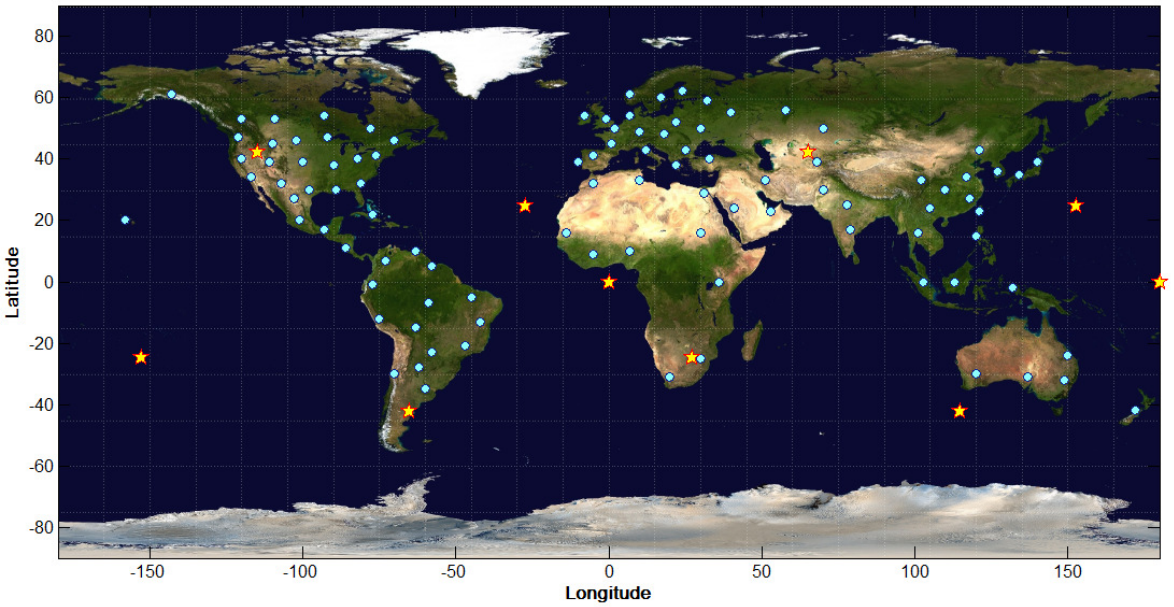


Figure 6.6. Locations of HAPs and initial location of satellites (System 1 – ICO)

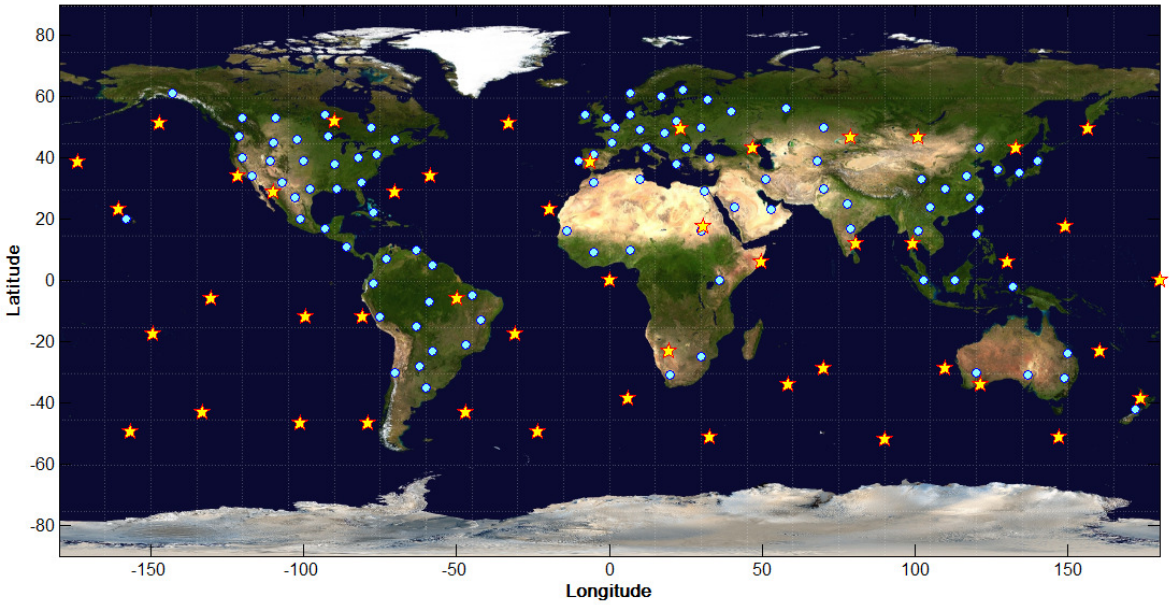


Figure 6.7. Locations of HAPs and initial location of satellites (System 2 – Globalstar)

Table 6.1. Characteristics of ICO and Globalstar Satellite Constellations

	ICO	Globalstar
Number of satellites ($N_P \times N_S$)	10 (2 × 5)	48 (6 × 8)
Altitude (h_S)	10355 km	1410 km
Period (T_{sat})	6 hours	2 hours
Inclination angle (α_I)	45˚	52˚
Type	Inclined 2π-constellation	Inclined 2π-constellation
Phase difference ($\Delta\phi$)	0˚	$360°/N_P N_S = 7.5°$

We set Δt to 1 minute and simulate the system for a system period (T_S) after a 2 hours of warm-up period. Theoretically, since HAPs are located above the cloud layer, optical data link can be established at an elevation angle even below the horizon (-2˚) [74]. However, for low elevation angles, some challenges must be surmounted. As the optical beam sinks toward the horizon and below, one expects longer propagation distance, stronger atmospheric attenuation, stronger wavefront distortions and scintillation, larger Doppler shift, possible obstructions of the beam due to HAP geometry, and possible interference of the Sun on the satellite terminal [75]. Therefore, we first test the performance of the system for ε_{min} value of -2˚. Then we compare the results for a system that doesn't allow elevation angles below 10˚ (ε_{min}=10˚).

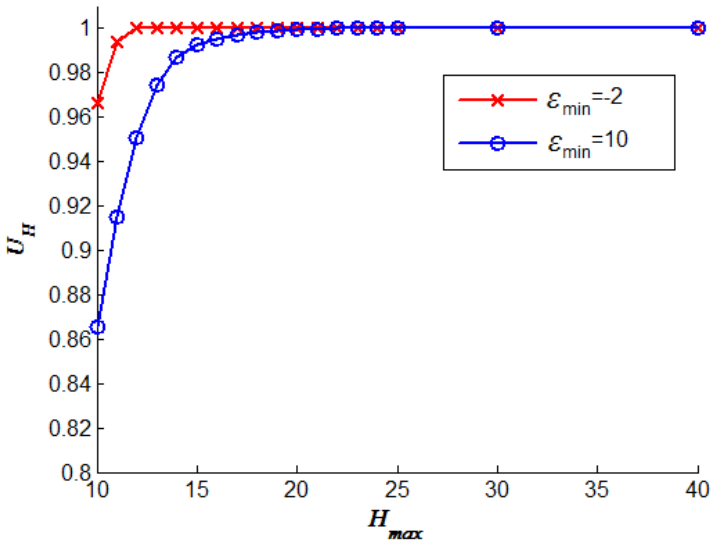


Figure 6.8. Utilization of HAPs for different H_{max} values. (System-1)

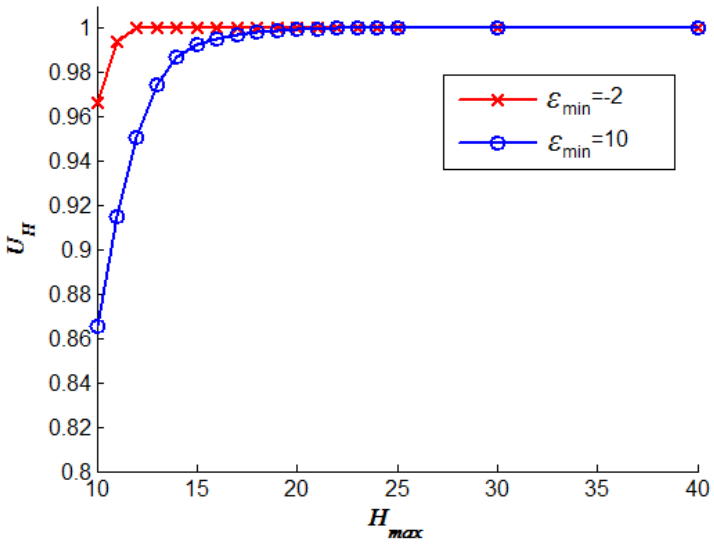


Figure 6.9. Average of elevation angles for different $H_{\max}$ values. (System-1)

Figure 6.8 illustrates the obtained U_H values for different $H_{\max}$ values, for System-1. Recall that the obtained values are the maximum possible utilization values under the given conditions. For low $H_{\max}$ values it is not possible to serve all of the HAPs due to satellite resource limitations. For $\varepsilon_{\min}$=10˚, degradation in utilization is more, since less number of satellites are visible to HAPs. For large $H_{\max}$ values, full utilization can always be achieved. To enable full utilization of HAPs in the scenario, $H_{\max}$ should be at least 12 and 22 for $\varepsilon_{\min}$=-2˚ and $\varepsilon_{\min}$=10˚, respectively.

Figure 6.9 illustrates average of elevation angles between satellites and HAPs that are linked with each other. For low $H_{\max}$ values, average elevation angle is lower for the system with $\varepsilon_{\min}$=-2˚. This is due to the fact that, in that system, satellite-HAP pairs with very low elevation angles are linked in order to increase HAP utilization. As H_{max} exceeds a certain value, A_{avg} becomes independent of the minimum elevation angle, and for sufficiently large $H_{\max}$ values (larger than 30 in the scenario), it reaches to its maximum possible value (around 53˚ in the scenario).

Figure 6.10 and Figure 6.11 illustrate the simulation results for System-2. Obtained results show that performance behavior of System-2 with respect to changing $H_{\max}$ values is similar to the performance behavior of System-1. Since there are more satellites to serve

same number of HAPs, smaller H_{max} value is sufficient for full utilization of HAPs (it should be at least 4 and 6 for ε_{min}=-2° and ε_{min}=10°, respectively). On the other hand, elevation angle values are lower in System-2 comparing to the System-1. For example, maximum possible value of A_{avg} is approximately 42.5° in System-2, which is more than 10° lower than the maximum A_{avg} value for System-1. This is related to the altitudes and number of satellites in the system. Satellites with high altitudes have larger coverage areas and offer higher elevation angles according to 6.4. As the altitudes of satellites are decreased with a given ratio, the number of satellites should be sufficiently increased in order to acquire same coverage and elevation angle values. Although System-2 has more satellites than System-1, we can say that it does not offer higher coverage and elevation angle values due to low altitudes of satellites, according to the obtained results. Note that, although smaller H_{max} values are sufficient to achieve full utilization in System 2, due to its lower coverage, total number of optical transmitters/receivers needed in the whole system is not lower than System-1.

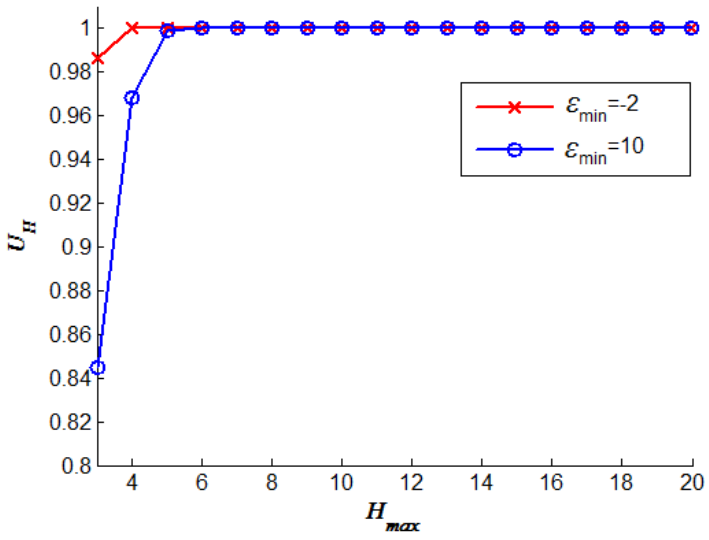


Figure 6.10. Utilization of HAPs for different H_{max} values. (System-2)

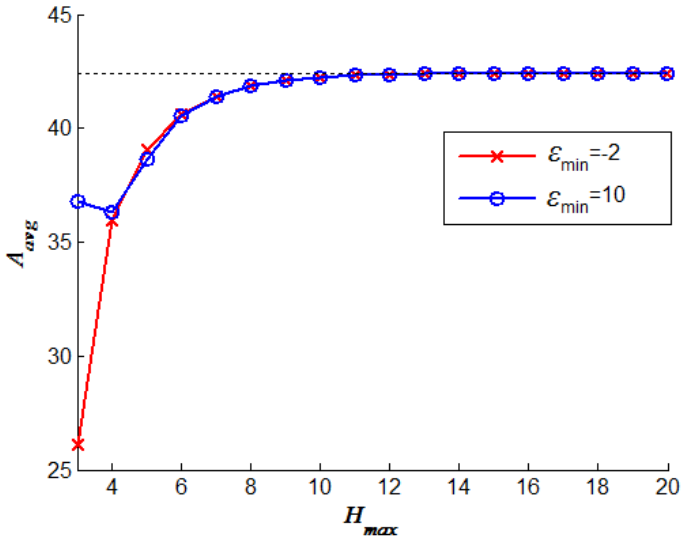


Figure 6.11. Average of elevation angles for different H_{max} values. (System-2)

Next, we test the average of link duration times. As we mentioned in the previous section, matching of satellites and HAPs that maximize the A_{avg} value may result in frequent switching of the optical links. Therefore, we proposed to favor existing links with a particular amount γ as described in Section 6.3.3. Figure 6.12 illustrates the average of link duration times, and Figure 6.13 illustrates the A_{avg} values with changing γ value, for System-1. If γ is set to zero, we never care about switching of links and get better A_{avg} value, but average link duration time is quite low. When we slightly increase γ value, link duration times effectively increase with a slight decrease in A_{avg} value. For example, for H_{max}=12, increasing γ from zero to one yields approximately 18% increase in link duration times with only 0.02% decrease in A_{avg}. Further increasing γ, gap between gain and loss gets closer. After a point, average link duration time reaches to a saturation value and do not increase with the γ value. For example increasing γ from 40 to 90, results in very slight increase in average link duration time, but significantly decreases A_{avg} value as shown in Figure 6.13. Considering these issues, an optimal value for γ could be chosen depending on the system objectives and requirements of the mission-critical applications.

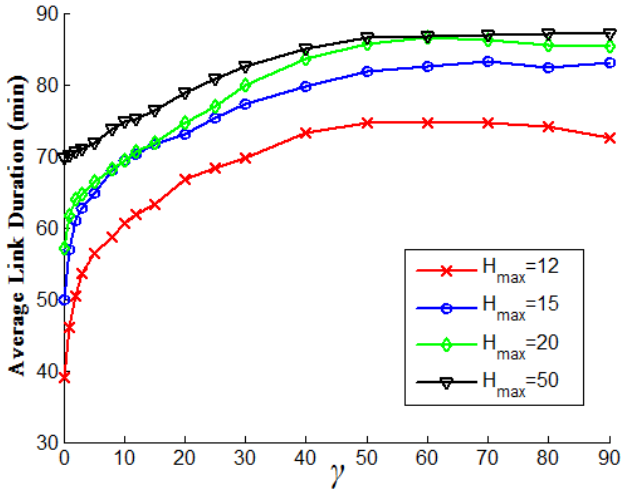


Figure 6.12. Average of link duration times for various γ and H_{max} values (System-1, ε_{min}=-2).

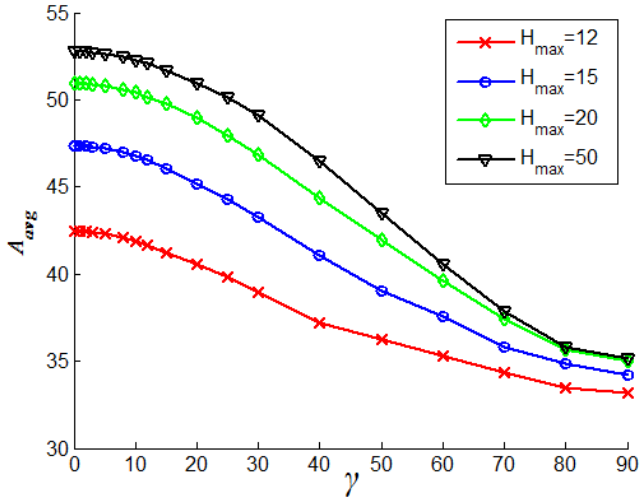


Figure 6.13. Average of elevation angles for various γ and H_{max} values (System-1, ε_{min}=-2).

Let us define gain and loss functions with respect to γ, $G(\gamma)$ and $L(\gamma)$, as follows:

$$G(\gamma) = \frac{LD(\gamma) - LD(0)}{LD(0)} \tag{6.17}$$

$$L(\gamma) = \frac{A_{\text{avg}}(0) - A_{\text{avg}}(\gamma)}{A_{\text{avg}}(0)} \tag{6.18}$$

where $LD(\gamma)$ denotes average link duration time, and $A_{avg}(\gamma)$ denotes A_{avg} value with respect to γ. $G(\gamma)$ represents ratio of gain obtained in link duration time, when we favor existing links by γ, and $L(\gamma)$ represents ratio of loss encountered in average elevation angles. Now, according to gain and loss functions, we define *net gain function* (*NGF*) as follows:

$$NGF(\gamma) = \eta \cdot G(\gamma) - (1-\eta) \cdot L(\gamma) \tag{6.19}$$

η is a value between zero and one and should be selected according to the objectives of the system. If minimizing the switching cost is more important than elevation angle, then η value should be closed to one, and if increasing the elevation angle is more important, then it should be closed to zero. Optimal γ value is the one that maximizes $NGF(\gamma)$ according to the selected η value. Figure 6.14 illustrates the obtained $NGF(\gamma)$ values with respect to different η values, for H_{max}=12. Maximum $NGF(\gamma)$ value obtained for each η value is marked with circle. In the considered scenario (System-1, H_{max}=12, ε_{min}=-1), best γ values for η = 0,8 and η = 0,2 are around 50 and 20, respectively. In general, one should decide for the most appropriate η value depending on the system objectives and requirements of the mission-critical applications, and multiple runs of optimization process could be performed for different γ values in order to converge to the maximum $NGF(\gamma)$ value.

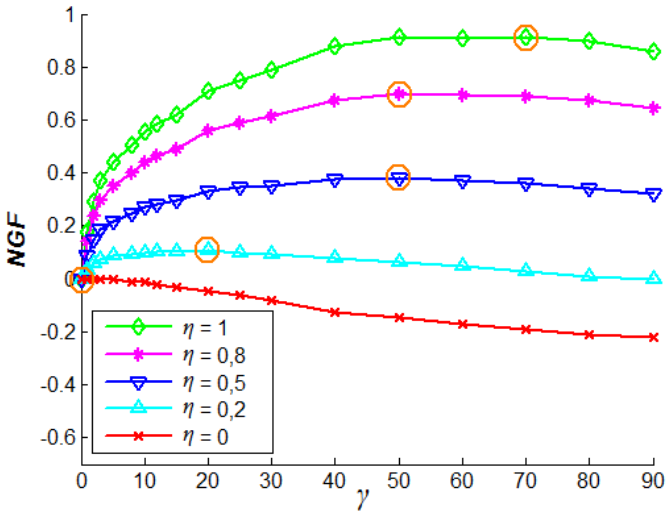


Figure 6.14. Net gain function for various η values (System-1, H_{max}=12, ε_{min}=-2)

6.5. Summary

In this chapter, we consider NGEO mobile satellite systems integrated with HAPs. HAPs are very suitable for most mission-critical applications, because they may be located over sensitive areas, receive and generate large amount of mission-critical data, and then transmit to Mission Control Centers (MCC) via satellites using high capacity free-space optical links. In this system, since there exists more than one satellite in line of sight, we investigate the problem of deciding on which satellite-HAP pairs to establish optical link. This problem is very crucial for acquiring maximum system performance and full system availability in the mission-critical network. In this context, we consider the minimum elevation angle constraint and the limit on the satellite resources. We propose a problem formulation for maximizing the utilization of HAPs, as well as maximizing the average elevation angle between HAPs and satellites. Moreover, we also propose a method for avoiding frequent switching of optical links, which is an expensive task. We come up with a polynomial-time solution approach for the formulated optimization problem using a combinatorial graph algorithm. We perform the proposed optimization in sample system scenarios and provide the resulting utilization, average elevation angle, and average link duration time values for various system parameters. Simulation results show the effect of resource restrictions on the system performance, and point out the minimum number of optical transmitters/receivers needed for enabling full utilization of HAPs and for achieving maximum possible value of average elevation angle. Moreover, we observe that proposed method for avoiding frequent switching is very effective, and with an appropriate selection of system parameters, significant performance improvement can be achieved.

7. CONCLUDING REMARKS

This thesis clearly identifies the features of NGEO satellite networks that differs them from the terrestrial systems. For efficient networking, these features should be considered, and challenges due to mobility and resource limitations of satellite nodes should be handled properly. We classified the networking challenges in NGEO satellite networks, and provided a thorough survey about the routing and network mobility management strategies for overcoming these challenges. This thesis pointed out open issues in this context, and presented novel methods for efficient networking in NGEO satellite networks. In particular, geometrical properties and dynamics of regular satellite constellation topologies are considered, and novel routing and network mobility management techniques are developed.

Firstly, regarding the geometrical properties of NGEO satellite constellation topologies, a traffic-sensitive priority-based adaptive routing (PAR) algorithm is introduced for use in NGEO satellite networks. In the PAR algorithm, rather than setting the route in the terrestrial nodes or in a single satellite node, the route is set-up by making decision of sending packet from which outgoing link, at each hop. The decision criterion depends on a priority mechanism, which favors links that are less utilized. By this way, more utilization of links may be provided. Enhanced PAR (ePAR) algorithm is also proposed and analyzed in order to further enhance the routing algorithm for providing channeling of packets with same source-destination pairs to same links. The proposed priority mechanism do not have any signaling overhead, and based on extensive set of simulations, it is shown to be promising for use in NGEO satellite networks. Further, deflection enabled PAR algorithm (DEPAR) is proposed as an extension of PAR algorithm. DEPAR deflects the packets to longer routes when the outgoing links in shortest paths are not available. Simulation results show that proposed deflection routing approach is promising for low traffic loads, but it fails to improve performance for high traffic loads. Including traffic load sensitivity to the deflection mechanism would be an interesting subject of a future study.

Second major contribution of this thesis is related to handling mobility of satellites. Mobility of satellites is a major challenge especially for connection-oriented data communication in next generation satellite networks. VN concept is proposed for mobility handling, however it requires one-to-one correspondence between physical satellites and virtual nodes, resulting in the reduced system availability. In this thesis, we investigated more general virtual topology characteristics for satellite systems with Earth-fixed footprints, where more than one satellite can serve for the same footprint area. We provided formal model for proposed multi-state virtual network (MSVN) topology and pointed out its possible contributions to the overall system availability. We investigated potential handover mechanisms for VN-based and MSVN-based satellite systems, and proposed efficient handover algorithms, namely VN-HO, MSVN-SHO and MSVN-SSHO. Despite a marginal increase in the cost, MSVN-based systems offer handover algorithms that are faster and smoother than VN-HO. Moreover, an optimal beam management technique is proposed to show that system availability and performance of MSVN-based systems can be significantly increased by directing beams to denser areas. These construct significant benefits of MSVN-based satellite systems over conventional VN-based satellite systems.

A possible future work in this scope is to develop an MPLS-based Earth-fixed system for both VN and MSVN topologies. Previously, it is stated that developing highly efficient rerouting mechanisms is the most crucial issue for employing MPLS in satellite constellations [48]. Since Earth-fixed satellite systems significantly simplify rerouting issues, they are very appropriate for employing MPLS.

Finally, integration of mobile NGEO satellite systems with High Altitude Platforms (HAPs) is investigated. HAPs and satellites can communicate via high capacity free space optical links, however resource restrictions and elevation angles should be taken into account for optimal integration. In order to maximize the utilization of HAPs, as well as the average elevation angle between HAPs and satellites, problem of optimal assignment of satellites to HAPs is formulated and solved. Moreover, a technique is proposed for avoiding frequent switching of optical links. Simulation results show the effects of resource limitations to the system performance, and point out minimum amount of satellite resources required for full utilization of HAPs and for achieving maximum performance.

Moreover, it is observed that proposed method for avoiding frequent switching is very effective, and with an appropriate selection of system parameters, significant performance improvement can be achieved. It is concluded that considered integrated scenario and optimal integration techniques have a great potential to satisfy the needs of emerging mission-critical applications. Multicast routing in the considered integrated scenario could be an interesting future work.

REFERENCES

1. Wood, L., *Internetworking with Satellite Constellations*, Ph.D. Thesis, University of Surrey, June 2001.

2. Jamalipour, A., *Low Earth Orbital Satellites for Personal Communication Networks*, Norwood, MA: Artech House, 1998.

3. Ghedia, L., K. Smith, and G. Titzer, "Satellite PCN - The ICO System," *International Journal of Satellite Communications*, Vol. 17, pp. 273-289, July/Aug., 1999.

4. Schindall, J., "Concept and Implementation of the Globalstar Mobile Satellite System," *Proceedings 4th International Mobile Satellite Conference (IMSC '95)*, pp. A11–A16, Ottawa, Canada, June 1995.

5. Leopold, R. J. and A. Miller, "The Iridium Communication System," *IEEE Potentials*, Vol. 12, pp. 6-9, Apr. 1993.

6. Sturza, M. A., "Architecture of Teledesic Satellite System," *Proceedings of the 4th International Mobile Satellite Conference (IMSC'95)*, pp. 212-218, Ottawa, Canada, June 1995.

7. Ferreira, A., J. Galtier, and P. Penna, "Topological Design, Routing and Hand-over in Satellite Networks", in I. Stojmenovic, *Handbook of Wireless Networks and Mobile Computing*, pp. 473-507, John Wiley and Sons Ltd., London, 2002.

8. Walker, J. G., "Circular Orbit Patterns Providing Continuous Whole Earth Coverage," Tech. Rep. 70211 (UDC 629.195:521.6), Royal Aircraft Establishment, UK, Nov. 1970.

9. Restrepo J., G. Maral, "Cellular Geometry for World-wide Coverage by Non-GEO Satellites using 'Earth-fixed Cell' Technique," *Space Communications*, Vol. 14, pp. 179-189, 1996.

10. Alagöz, F., Ö. Korçak, A. Jamalipour, "Exploring the Routing Strategies in Next-Generation Satellite Networks", *IEEE Wireless Communications Magazine*, Vol. 14, No. 3, June 2007.

11. Korçak, Ö., F. Alagöz, "Priority-based Adaptive Shortest Path Routing in Next-Generation LEO-Satellite Networks," *Proceedings of 23rd AIAA International Communications Satellite Systems Conference (ICSSC)*, Rome, Italy, September 2005.

12. Korçak, Ö., F.Alagöz, A.Jamalipour, "Priority-based Adaptive Routing in NGEO Satellite Networks," *International Journal of Communication Systems*, Vol. 20, No. 3, pp. 313-333, March 2007.

13. Korçak, Ö., F.Alagöz, "Deflection Routing over Prioritized Intersatellite Links in LEO Satellite Networks," *Proceedings of 2nd IEEE International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA)*, pp. 2485-2490, Damascus, Syria, April 2006.

14. Korçak, Ö., F.Alagöz, "Multi-state Virtual Network Architecture for Next Generation Satellite Networks," *Proceedings of IEEE GLOBECOM*, pp. 5031-5036, Washington D.C., Nov. 2007.

15. Korçak, Ö., F.Alagöz, "Virtual Topology Dynamics and Handover Mechanisms in Earth-Fixed Satellite Systems," *Computer Networks*, Vol. 53, No. 9, pp. 1497-1511, Jun. 2009.

16. Korçak, Ö., F.Alagöz, "Link-layer Handover in Earth-fixed LEO Satellite Systems," *Proceedings of IEEE ICC*, Dresden, Germany, Jun. 2009.

17. Korçak, Ö., F.Alagöz, "Optimal Beam Management in Earth-fixed Satellite Systems," *Proceedings of International Workshop on Satellite and Space Communications (IWSSC)*, pp. 43-47, Salzburg, Austria, Sept. 2007.

18. Korçak, Ö., F.Alagöz, "Efficient Networking in an Integrated HAP and Mobile Satellite System with Optical Links", *Proceedings of IEEE/IFIP 6th International Conference on Wireless and Optical Comminications Networks (WOCN)*, Cairo, Egypt, Apr. 2009.

19. Coltun, R. and V. Fuller, "The OSPF NSSA Option," *Network Working Group*, RFC 1587, Mar. 1994.

20. Meyer, G. and S. Sherry, "Triggered Extensions to RIP to Support Demand Circuits," *Network Working Group*, RFC 2091, Jan. 1997.

21. Chang, H. S., B. W. Kim, C. G. Lee, S. L. Min, Y. Choi, H. S. Yang, D. N. Kim and C. S. Kim, "Performance Comparison of Optimal Routing and Dynamic Routing in Low-Earth Orbit Satellite Networks," *Proceedings of VTC'96*, Atlanta, GA, 1996.

22. Werner, M., "A Dynamic Routing Concept for ATM Based Satellite Personal Communication Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, pp. 1636–48, Oct. 1997.

23. Gounder V. V., R. Prakash, H. Abu-Amara, "Routing in LEO-based Satellite Networks," *Proceedings of IEEE Emerging Technologies Symposium on Wireless on Wireless Communications and Systems,* Richardson, USA, April 1999.

24. Mauger, R., C. Rosenberg, "QoS Guarantees for Multimedia Services on a TDMA-based Satellite Network," *IEEE Communications Magazine*, vol. 35(7), pp. 56-65, 1997.

25. Wood, L., A. Clerget, I. Andrikopoulos, G. Pavlou, and W. Dabbous, "IP Routing Issues in Satellite Constellation Networks," *International Journal of Satellite Networks*, Vol. 19, No. 1, pp. 69-92, Jan/Feb 2001.

26. Ekici, E., I. F. Akyildiz, and M. D. Bender, "A Distributed Routing Algorithm for Datagram Traffic in LEO Satellite Networks," *IEEE/ACM Transactions on Networking*, Vol. 9, No. 2, pp. 137-147, 2001.

27. Ekici, E., I. F. Akyildiz, M. D. Bender, "A Multicast Routing Algorithm for LEO Satellite IP Networks", *IEEE/ACM Transactions on Networking*, Vol. 10, No. 2, pp. 183-192, April 2002.

28. Akyildiz, I. F., E. Ekici, M. D. Bender, "MLSR: A novel routing algorithm for multi-layered satellite IP networks," *IEEE/ACM Transactions on Networking*, Vol. 10, No. 3, pp. 411-424, 2002.

29. Bayhan, S., G. Gür, F. Alagöz, "VoIP Performance in Multilayered Satellite IP Networks with OBP", *International Journal of Communication Systems*, 2007, Vol. 20, No. 12, pp. 1367-1389, 2007.

30. Dash, D. S., A. Durresi, R. Jain, “Routing of VoIP traffic in Multilayered Satellite Networks,” *Proceedings of SPIE ITCOMM* 2003, pp. 65-75, Orlando, Florida, vol. 5244.

31. Werner, M., C. Delucchi, H. J. Vogel, G. Maral, and J. J. De Ridder, "ATM-Based Routing in LEO-MEO Satellite Networks with Intersatellite Links", *IEEE Journal on Selected areas in Communications*, Vol. 15, No. 1, pp. 69-82, Jan. 1997.

32. Erçetin, Ö., S. Krishnamurthy, S. K. Dao and L.Tassiulas, “A Predictive QoS Routing Scheme for Broadband Low Earth Orbit Satellite Networks,” *Proceedings of Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1064-1074, London, UK, 2000.

33. Jukan, A., N.H. Nguyen, H.R. van As, “An Approach to QoS-based Routing for LEO Satellite Networks”, *Proceedings of IEEE International Conference on Communications Technology*, pp. 922-929, Beijing, China, 2000.

34. Chen, J., and A. Jamalipour, “An Adaptive Path Routing Scheme for Satellite IP Networks”, *International Journal of Communication Systems*, Vol. 16, No. 1, pp. 5–21, February 2003.

35. Uzunalioglu, H., M. D. Bender, and I. F. Akyildiz, “A Routing Algorithm for LEO Satellite Networks with Dynamic Connectivity,” *ACM-Baltzer Journal of Wireless Networks (WINET)*, Vol. 6, No. 3, pp. 181-190, 2000.

36. Uzunalioğlu, H., I. F. Akyildiz, Y. Yesha, W. Yen, “Footprint Handover Rerouting Protocol for Low Earth Orbit Satellite Networks,” *ACM-Baltzer J. Wireless Networks*, Vol. 5, No. 5, pp 327-337, Nov. 1999.

37. Sun, J., E. Modiano, "Routing Strategies for Maximizing Throughput in LEO Satellite Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 22, pp. 273-286, 2004.

38. Henderson, T. H., and R.H. Katz, "On Distributed and Geographic-based Packet Routing for LEO Satellite Networks", *Proceedings of IEEE Globecom*, pp. 1119-1123, San Francisco, CA, USA, Dec. 2000.

39. Franck, L., and G. Maral, "Static and Adaptive Routing in ISL Networks from a Constellation Perspective", *The International Journal of Satellite Communications*, Vol. 6, No. 20, pp 455-475, 2002.

40. Papapetrou E. and F-N. Pavlidou, "A Proposal of Optimal Routing Techniques for Non-GEO Satellite Systems," *International Journal of Wireless Information Networks*, Vol. 8, No. 2, pp. 75-83, 2001.

41. Bertsekas, D., L. Gallager , *Data Networks*, Prentice-Hall, New Jersey, 2nd edition 1992.

42. Taleb, T., D. Mashimo, A. Jamalipour, N. Kato, and Y. Nemoto, "ELB: An Explicit Load Balancing Routing Protocol for Multi-hop NGEO Satellite Constellations," *Proceedings of IEEE Globecom*, pp. 1-5, San Francisco, USA, Nov. 2006.

43. Sigel, E., B. Denby, and S. Le Hegarat-Mascle, "Application of Ant Colony Optimization to Adaptive Routing in a LEO Telecommunications Satellite Network", *Annals of Telecommunications*, Vol.57, No. 5-6, pp. 520-539, May-June 2002.

44. Lee, J., and S. Kang, "Satellite Over Satellite (SOS) Network: A Novel Architecture for Satellite Network," *Proceedings of IEEE INFOCOM*, Vol. 1, pp. 315-321, Tel-Aviv, Mar. 2000.

45. Chen, C. and E. Ekici, "A Routing Protocol for Hierarchical LEO/MEO Satellite IP Networks," *Wireless Networks*, Vol. 11, No. 4, pp. 507-523, July 2005.

46. Kandus, G., A. Svigelj, and M. Mohorcic, "The Impact of Different Scheduling Policies on Traffic Class Dependent Routing in Intersatellite Link Networks," *International Journal of Satellite Commun.*, Vol. 22, pp. 533-546, 2004.

47. Kota, S. L., K. Pahlavan, P. Leppanen, *Broadband Satellite Communications for Internet Access*, Kluwer Academic Publishers, 2004.

48. Donner, D., M. Berioli, and M. Werner, "MPLS-based Satellite Constellation Networks", *IEEE Journal on Selected areas in Communications*, Vol. 22, No. 3, pp. 438-448, Apr. 2004.

49. Deering, S. E., and D. R. Cheriton, "Multicast Routing in Datagram Internetworks and Extended LANs," *ACM Trans. Comput. Syst.*, Vol. 8, pp. 85–110, May 1983.

50. Waitzman, D., C. Partridge, and S. Deering, *Distance Vector Multicast Routing Protocol*, RFC 1075, Nov. 1988.

51. Moy, J., "Multicast Routing Extensions to OSPF," *Commun. ACM*, Vol. 37, pp. 61–66, Aug. 1994.

52. Korçak, Ö., I. Kaya, H. G. Çalıklı, and F. Alagöz, "Performance Evaluation of Adaptive and Static Routing Algorithms and Contention Resolution Techniques in

LEO Satellite Constellations," *Proceedings of IEEE Recent Advances in Space Technologies RAST*, pp. 207-212, Istanbul, Turkey, June 2005.

53. Küçükateş, R., and C. Ersoy, "Minimum Flow Maximum Residual Routing in LEO Satellite Networks Routing Set," *Wireless Networks*, Vol. 14, No. 4, pp. 501-517, September, 2008.

54. Jianjun, B., L. Xicheng, L. Zexin, and P. Wei, "Compact Explicit Multi-path Routing for Low Earth Orbit Satellite Networks," *Proceedings of IEEE Workshop on High Performance Switching and Routing*, pp. 386-390, Hong Kong, P.R.China, May 2005.

55. Voilet, M. D., *The Development and Application of a Cost per Minute Metric of the Evaluation of Mobile Satellite Systems in a Limited-growth Voice Communications Market*, Master's thesis, MIT, 1995.

56. "Distribution of Top-level Domain Names by Host Count Jan 2005", Internet System Consortium, http://www.isc.org, [cited: May 2005].

57. Perdigues, J., M. Werner, and K. Karafolas, "Methodology for Traffic Analysis and ISL Capacity Dimensioning in Broadband Satellite Constellations Using Optical WDM Networking," *Proceedings of 19th AIAA International Communication Satellite Systems Conference (ICSSC'01)*, Toulouse, France, April, 2001.

58. Chang, H. S., B. W. Kim, C. G. Lee, Y. Choi, S. L. Min, H. S. Yang, and C. S. Kim, "Topological Design and Routing for Low-Earth-Orbit Satellite Networks," *Proceedings of IEEE Globecom*, pp. 529-535, Singapore, 1995.

59. Chowdhury, P. K., M. Atiquzzaman, and W. Ivanvic, "Handover Schemes in Satellite Networks: State-of-the-Art and Future Research Directions," *IEEE Communications Surveys & Tutorials*, Vol. 8, No. 4, August 2006.

60. Patterson, D. and M. Sturza, U.S. Patent No. 5408237: "Earth-Fixed Cell Beam Management for Satellite Communication System", 1995.

61. Sarikaya, B. and M. Tasaki, "Supporting Node Mobility Using IPv6 in a LEO-satellite Network," *International Journal of Satellite Communications*, Vol. 19, No. 5, pp. 481-498, 2001.

62. Ahuja, R. K., T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, 1993.

63. Jamalipour, A. and T. Tung, "The Role of Satellites in Global IT: Trends and Implications", *IEEE Personal Comm.*, Vol. 8, No. 3, pp 5-11, 2001.

64. Karapantazis, S. and F-N. Pavlidou, "Broadband Communications via High Altitude latforms: A Survey", *IEEE Communications Surveys & Tutorials*, Vol. 7, No. 1, pp. 2-31, 2005.

65. Cianca, E. et *al*., "Integrated Satellite-HAP Systems," *IEEE Radio Communications*, Vol. 43, No. 12, pp. 33-39, 2005.

66. Gace, P., et *al*., "An Integrated Satellite-HAP-Terrestrial System Architecture: Resource Allocation and Traffic Management Issues," *Proceedings of IEEE VTC*, Milan, Italy, May, 2004.

67. Karapantazis, S. and F-N. Pavlidou, "The Role of High Altitude Platforms in Beyond 3G Networks," *IEEE Wireless Communications*, Vol. 12, No. 6, pp. 33-41, 2005.

68. Knapek, M. et *al.*, "Optical High-capacity Satellite Downlinks via High Altitude Platform Relays", *SPIE Free Space Laser Communications VII*, San Diego, 2007.

69. Antonini, M. et *al.*, "Feasibility Analysis of a HAP-LEO Optical Link for Data Relay Purposes", *Proceedings of IEEE Aerospace Conference*, Big Sky, MT, USA, March, 2006.

70. Carrozzo, V. and G. Parca, "Hybrid Network Based on Intersatellite Communication Links and WDM Technology", *Wireless Communications and Mobile Computing*, DOI: 10.1002/wcm.675, 2008.

71. Farserotu, J., G. Kotrotsios, I. Kjelberg and A. Prasad, "Scalable, Hybrid Optical-RF Wireless Communication System for Broadband and Multimedia Service to Fixed and Mobile Users", *Wireless Personal Communications*, Vol. 24, pp. 327-339, 2003.

72. Kuhn, N. W., "The Hungarian Method for the Assignment Problem", *Naval Research Logistics Quarterly*, Vol. 2, pp. 83-97, 1955.

73. Carpaneto, G., S. Martello, and P. Toth, "Algorithms and Codes for the Assignment Problem", *Annals of Operations Research*, Vol. 13, pp. 193-223, 1988.

74. Giggenbach, D., J. Horwath, and B. Epple, "Optical Satellite Downlinks to Optical Ground Stations and High-Altitude Platforms," *IST Mobile and Wireless Communication Summit*, Budapest, July, 2007.

75. Perlot, N. et *al.*, "System Requirements for Optical HAP-Satellite Links", *Proceedings of 6th Int. Symposium on Comm. Systems, Networks and Digital Signal Processing (CSNDSP)*, pp. 72-76, Graz, Austria, July, 2008.